

Documentação da parte de Programação do Projeto Big Data - Módulo 8 - Inteli

Grupo Pérola Negra - Solução DataApp com Dashbord

Integrantes do Grupo:

- Ana Martire
- Eduardo Oliveira
- Keylla Oliveira
- Lucas Barbosa
- Nicollas Isaac
- Sophia Nóbrega

Sumário

- Documentação da parte de Programação do Projeto Big Data - Módulo 8 - Inteli
 - Grupo Pérola Negra - Solução DataApp com Dashbord
 - Integrantes do Grupo:
- Sumário
- 1. Data Product Canvas
 - Problema
 - Dados
 - Solução
 - Hipóteses
 - KPIs
 - Atores
 - Ações
 - Valores
 - Riscos
 - Performance/Impacto
- 2. Arquitetura Macro
 - 2.1. Componentes da Arquitetura
 - 2.2 UML de Componentes
 - Arquitetura Medallion
 - Camada de Bronze (Data Lake)
 - Camada de Prata (Transformação e Normalização)
 - Camada de Ouro (Data Warehouse)
 - Camada de Ródio (Visualização)
 - Ingestão
 - Transformação
 - Análise
 - Visualização
- 3. Processo de ETL

- ETL
 - Arquitetura e Fluxo do Pipeline de Dados
 - Resumo do Pipeline
 - 3.1 Análise de Dados
 - 3.1.1 Criação das Planilhas
 - 3.2 Cubo de Dados
- 4. Análise de Impacto Ético
 - Introdução
 - Impactos em Meio Ambiente e Sociedade
 - 4.1. Privacidade e Proteção de Dados
 - 4.2. Equidade e Justiça
 - 4.3. Transparência e Consentimento Informado
 - 4.4. Responsabilidade Social
 - 4.5. Viés e Discriminação
 - 4.6. Responsabilidade social
 - Conclusão
- 5. Streamlit e Infográfico
 - 5.1. Documentação do Streamlit
 - 5.1.1. Autenticação
 - 5.1.2. Dashboard
 - 5.2. Documentação dos Filtros
 - 5.3. Documentação do Infográfico
 - 5.4 Documentação dos Relatórios
- 6. Cobertura de Testes
 - 6.1. Objetivo
 - 6.2 Estrutura de Testes
 - 1. Testes para Processos ETL
 - Casos Testados
 - Ferramentas Utilizadas
 - 2. Testes para Views
 - Casos Testados
 - Integração com Streamlit
 - 3. Geração de Relatórios
 - Passos para Geração do Relatório
 - 6.3. Conexões no Streamlit para Testes
 - 6.4. Conclusão
- 7. Conclusões e Próximos Passos
 - 7.1. Conclusões Obtidas
 - 7.2. Próximos Passos
 - 7.3. Outras Perspectivas e Ideias Futuras
- 8. Anexos
 - Anexo I
- Termo de Uso e Política de Privacidade de Dados
 - 1. Disposições Gerais
 - 2. Objetivo da Coleta de Dados
 - 3. Tipos de Dados Coletados

- 4. Consentimento e Revogação
 - 4.1 Obtenção de Consentimento
 - 4.2 Revogação de Consentimento
- 5. Transparência e Acesso à Informação
- 6. Segurança e Armazenamento de Dados
- 7. Direitos dos Usuários
- 8. Auditoria e Conformidade
- 9. Canal de Atendimento ao Usuário
- 10. Disposições Finais
- 9. Automatização de Coleta
 - 9.1. Automatização do Data Ingestion
 - 9.2. Conclusão
- 10. Referências

1. Data Product Canvas

Dividido em 10 blocos (problema, solução, dados, hipóteses, atores, ações, KPIs, valores, riscos e performance/impacto), o Data Product Canvas é um framework desenvolvido para auxiliar no planejamento estratégico e execução de produtos de dados. Baseado na Metodologia Ágil/Lean, ele organiza as informações essenciais do projeto em um modelo visual, alinhando a visão de todos os stakeholders. Seu principal objetivo é proporcionar clareza sobre o propósito do projeto e facilitar a geração de um roadmap estruturado.

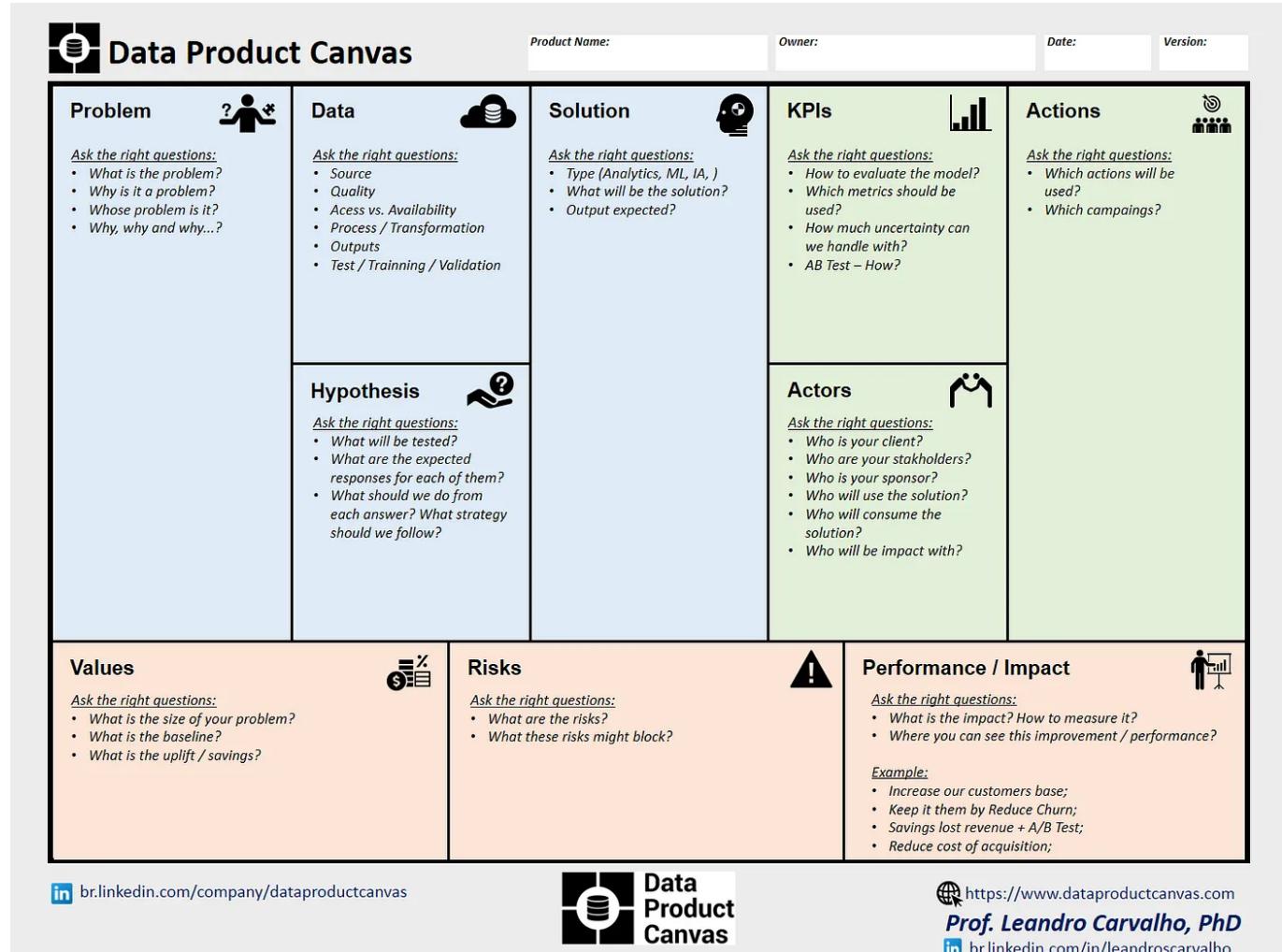
Cada bloco do Canvas é distribuído em 3 áreas de domínio:

1. **Visão do produto:** composta pelos blocos problema, solução, dados e hipóteses;
2. **Visão da estratégia:** composta pelos blocos atores, ações e KPIs;
3. **Visão do negócio:** composta pelos blocos valores, riscos e performance/impacto.

O uso do Data Product Canvas no projeto da CPTM destaca sua aplicação prática na resolução de problemas operacionais críticos. O framework foi utilizado para mapear as necessidades de eficiência e segurança no transporte público, priorizando o monitoramento automatizado de operações ferroviárias, como falhas captadas pelos sensores, fluxo de passageiros e sincronização de portas.

Além disso, a proposta de valor do produto está diretamente alinhada às demandas específicas da CPTM. O pipeline desenvolvido permite ações rápidas e corretivas, reduzindo o impacto das falhas e melhorando o atendimento ao cliente. Por exemplo, com notificações em tempo real, operadores e técnicos conseguem identificar e corrigir problemas antes que afetem o fluxo de passageiros, garantindo maior confiabilidade na operação.

A imagem abaixo apresenta o template do Data Product Canvas, demonstrando sua estrutura visual e organização em blocos. Cada bloco é preenchido com informações que detalham o Discovery necessário para uma compreensão única de cada parte do produto de dados que será desenvolvido.

Figura 1 - Template Data Product Canvas

Fonte: Leandro Carvalho (2024)

Em cada um é explorado em detalhes todo o Discovery necessário para que se tenha um entendimento único de cada parte do produto de dados que será desenvolvido. E cada domínio trata de uma área chave para o correto planejamento e desenvolvimento do produto, fornecendo uma visão 360 que vai da determinação do problema até a execução estratégica, passando pelo monitoramento de KPIs e mapeamento dos riscos. (Carvalho, 2024)

Figura 2 - DPC Caixa Preta



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Problema

O desafio principal está relacionado à incapacidade de detectar e reagir rapidamente a falhas ou anomalias operacionais captadas pela caixa preta dos trens. A análise manual dos dados pode ser lenta e propensa a erros, levando a ineficiências operacionais, falhas não identificadas e potenciais riscos à segurança.

Dados

O dataset inclui informações de diversas tabelas, principalmente relacionadas aos sensores dos trens e às operações da CPTM. Os mesmos são importantes para monitorar o comportamento do trem e identificar possíveis anomalias operacionais, como falhas de sensor, uso inadequado de portas e fluxo irregular de passageiros. Abaixo está o detalhamento das colunas de cada tabela:

Figura 3 - Schema SQL Tabelas

sua_tabela	dmo_anl_vw_tot_mov_periodo	dmo_anl_vw_tipo_embarque	dmo_anl_vw_intervalos_dia	users
int No datetime Open_Time datetime Closed_Time int Line_ID int Train_ID int StartStation_ID int Station_ID int NextStation_ID int EndStation_ID int Carriage_ID int Door_ID int IN int OUT int Command int SensorSts string filename float Door_Open_Duration int Hour category Time_Interval int Day_of_Week	int id_dt_hora_minuto int cod_bilh int cd_estac_bu datetime dt_validacao int total_validades category tipo_dia	int id_tipo_embarque category tx_movimento int cod_bilh int id_tipo_lancamento_fk category tx_lancamento	datetime dt_hora_minuto int id_dt_hora_minuto string hora_ini string hora_fim category tx_prefixo	category id category name category email float passwordhash int id_estacao int id_stacao category tx_nome int id_estacao_bu

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Abaixo encontra-se uma descrição dos dados de cada tabela fornecida pela CPTM. Vale ressaltar que tais interpretações foram realizadas pelos membros do grupo e não correspondem a uma descrição oficial dos dados.

- **Tabela: sua_tabela**

- **No** (int): Identificador único do registro.
- **Open_Time** (datetime): Horário de abertura da porta.
- **Closed_Time** (datetime): Horário de fechamento da porta.
- **Line_ID** (int): Identificador da linha do trem.
- **Train_ID** (int): Identificador do trem.
- **StartStation_ID** (int): Identificador da estação inicial.
- **Station_ID** (int): Identificador da estação atual.
- **NextStation_ID** (int): Identificador da próxima estação.
- **EndStation_ID** (int): Identificador da estação final.
- **Carriage_ID** (int): Identificador do vagão.
- **Door_ID** (int): Identificador da porta.
- **IN** (int): Quantidade de passageiros que entraram.
- **OUT** (int): Quantidade de passageiros que saíram.
- **Command** (int): Comando acionado para abertura ou fechamento.
- **SensorSts** (int): Status do sensor.
- **filename** (string): Nome do arquivo do log de dados.
- **Door_Open_Duration** (float): Duração da abertura da porta.
- **Hour** (int): Hora do evento.
- **Time_Interval** (category): Intervalo de tempo categorizado.
- **Day_of_Week** (int): Dia da semana.

- **Tabela: dmo_anl_vw_mov_periodo**

- **id_dt_hora_minuto** (int): Identificador da data e hora.
- **cod_bilh** (int): Código do bilhete.
- **cd_estac_bu** (int): Código da estação.
- **dt_validacao** (datetime): Data de validação.
- **total_validacoes** (int): Total de validações de bilhetes.
- **categoria** (category): Categoria de bilhete.

- **Tabela: dmo_anl_vw_tipo_embarque**

- **cd_tipo_embarque** (int): Código do tipo de embarque.
- **tx_movimento** (category): Tipo de movimento (entrada/saída).
- **cod_bilh** (int): Código do bilhete.
- **cd_tipo_lancamento_fk** (int): Código de lançamento.
- **tx_lancamento** (category): Tipo de lançamento.

- **Tabela: dmo_anl_vw_intervalos_dia**

- **dt_hora_minuto** (datetime): Data e hora.
- **id_dt_hora_minuto** (int): Identificador do minuto.
- **hora_ini** (string): Hora inicial do intervalo.

- **hora_fim** (string): Hora final do intervalo.
- **Tabela: dmo_anl_vw_estacoes**
 - **id_estacao** (int): Identificador da estação.
 - **tx_prefixo** (category): Prefixo da estação.
 - **tx_nome** (category): Nome da estação.
 - **cd_estacao_bu** (int): Código da estação.

- **Tabela: users**

- **id** (category): Identificador do usuário.
- **name** (category): Nome do usuário.
- **email** (category): E-mail do usuário.
- **passwordhash** (float): Hash da senha para autenticação.

Solução

Um pipeline automatizado para análise dos dados, no caso do grupo Pérola Negra, focado na caixa preta, com o objetivo de detectar e monitorar anomalias, além de otimizar as operações de embarque e desembarque. A solução pode envolver:

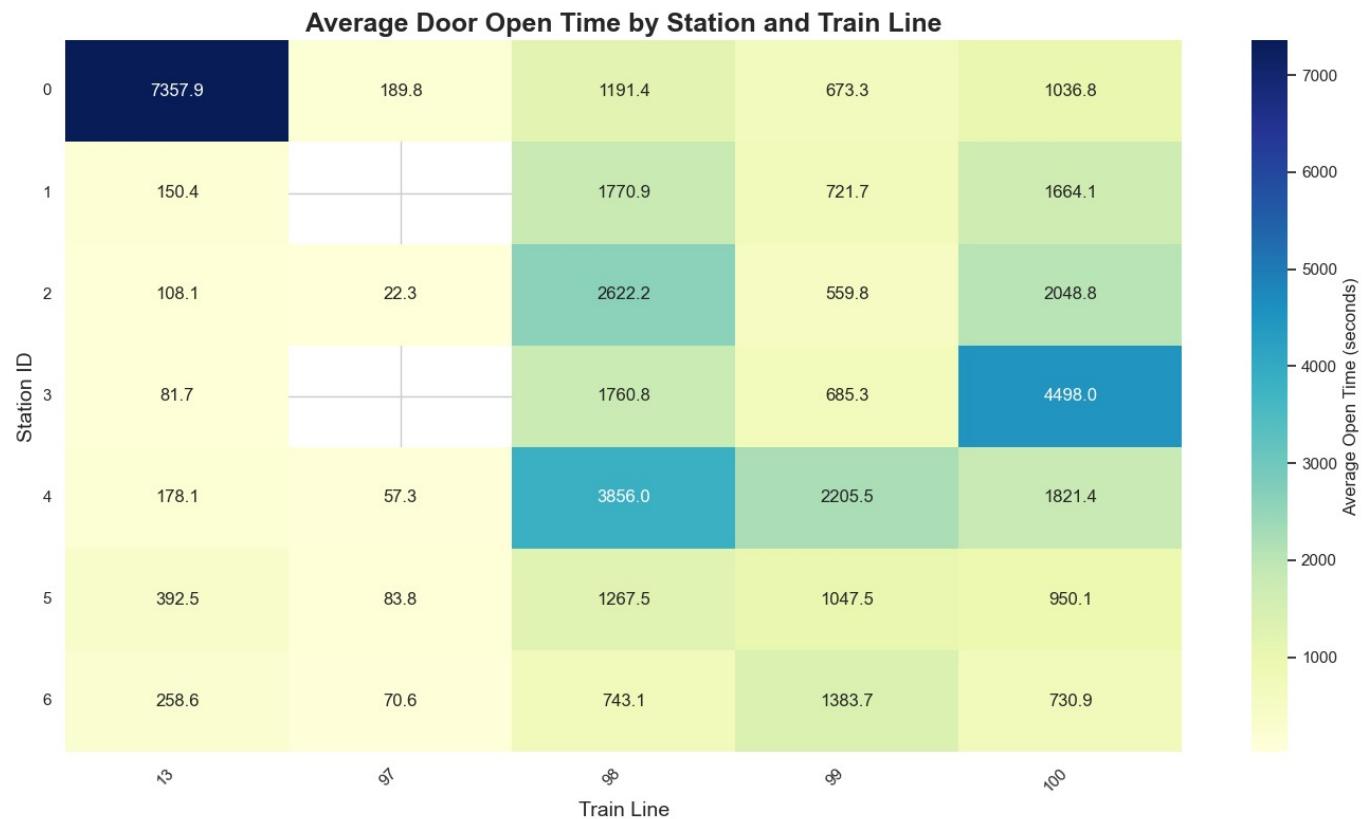
- Monitoramento contínuo de falhas nos sensores e comandos;
- Identificação de padrões operacionais críticos;
- Otimização da eficiência do fluxo de passageiros e alocação de frota com base nos dados históricos.

Hipóteses

- **H1:** A movimentação de passageiros (**IN** e **OUT**) varia significativamente com o horário do dia e a estação, sendo maior nas horas de pico e em estações centrais.
- **H2:** Certas portas são mais utilizadas, especialmente em vagões ou posições específicas do trem, o que pode impactar a eficiência operacional do embarque/desembarque.
- **H3:** Anomalias na sincronização das portas (diferença entre abertura e fechamento) impactam diretamente o tempo de parada nas estações.

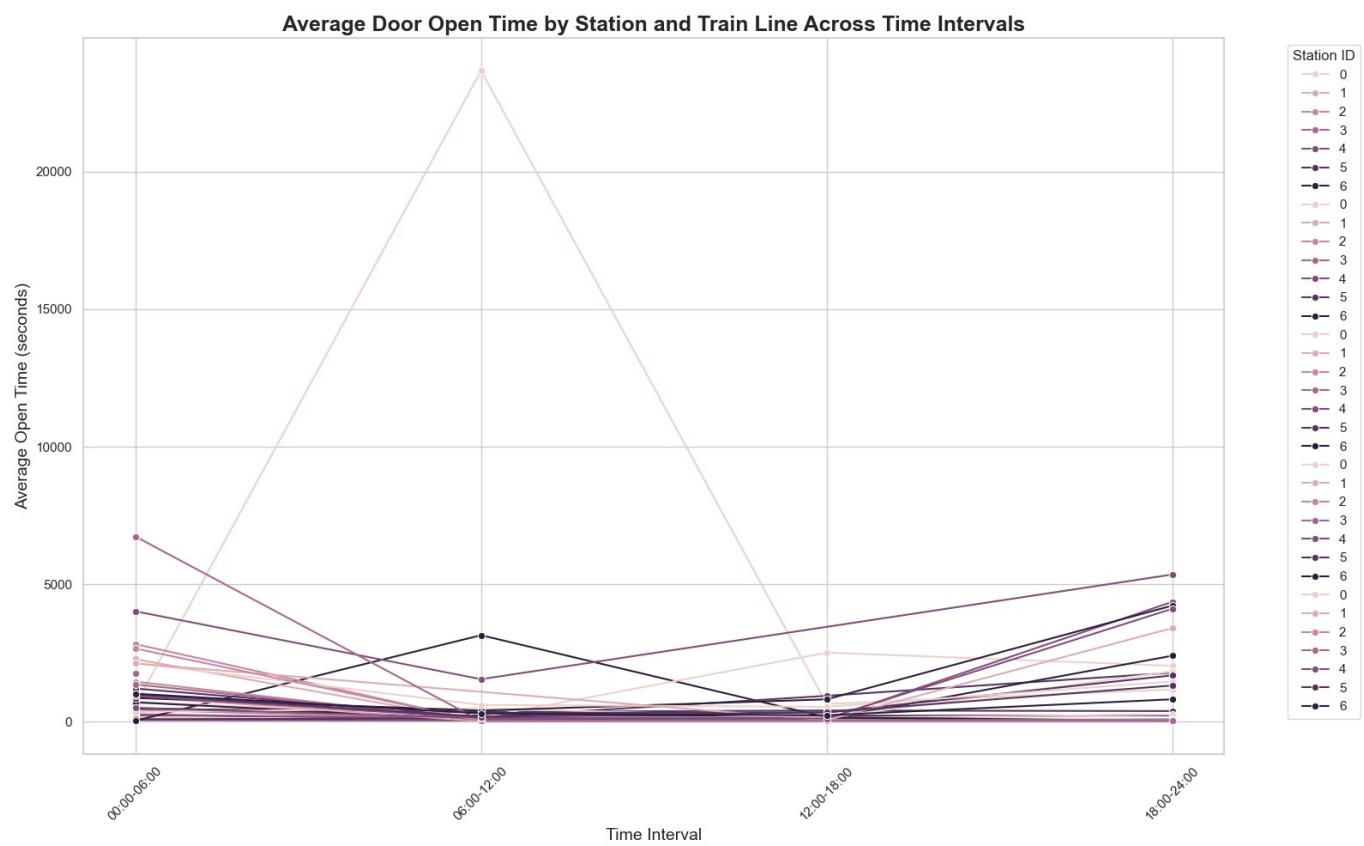
As hipóteses foram formuladas com base na análise exploratória realizada pela equipe. Identificamos padrões significativos através dos gráficos apresentados a seguir, que ilustram as correlações mencionadas:

Gráfico 4 - Média Abertura da Porta por Estação e Linha



Fonte: Leandro Carvalho (2024)

Gráfico 5 - Média Abertura da Porta por Estação e Linha com Intervalos de Tempo



Fonte: Leandro Carvalho (2024)

KPIs

1. **Monitoramento de embarque/desembarque por porta:** Avaliar a eficiência das portas durante o embarque e desembarque em diferentes horários e estações. Isso envolve medir o fluxo de passageiros (**IN** e **OUT**) por porta (**Door_ID**) e identificar padrões de uso. Este KPI ajuda a entender se certas portas ou vagões são subutilizados, o que pode guiar otimizações no layout dos trens e das plataformas.
2. **Controle da frota de carros por horários de pico:** Monitorar a quantidade de trens alocados em períodos de maior demanda (horários de pico) e garantir que a frota seja dimensionada adequadamente para reduzir a superlotação. A eficiência operacional pode ser medida verificando se o número de passageiros por trem é otimizado em horários críticos, melhorando a experiência do usuário e reduzindo custos.
3. **Análise de ocorrência de falhas com regressão linear:** Aplicar regressão linear para detectar padrões e prever falhas nos sensores e comandos. Este KPI busca identificar se há variáveis que indicam maior probabilidade de falhas (como o tipo de comando enviado ao trem, **Command**, e os problemas detectados nos sensores, **SensorSts**). O objetivo é antecipar e evitar falhas futuras, melhorando a segurança e a eficiência operacional.
4. **Aplicar melhorias no fluxo de passageiros em portas menos utilizadas:** Identificar padrões de movimentação de passageiros entre portas e vagões, otimizando tanto logística quanto layout do trem para melhorar o fluxo nas áreas menos movimentadas.
5. **Redução do tempo médio de inatividade dos trens:** Monitorar e reduzir o tempo necessário para diagnosticar e corrigir falhas operacionais, garantindo que os trens voltem à operação o mais rápido possível. Este KPI mede a eficiência do sistema em agilizar processos críticos.
6. **Aumento da pontualidade operacional:** Mensurar a quantidade de viagens que são concluídas no horário previsto, refletindo melhorias na eficiência geral do sistema. Esse KPI avalia diretamente o impacto das intervenções na confiabilidade do serviço.
7. **Diminuição do índice de falhas não detectadas:** Acompanhar a redução de falhas que passam despercebidas pelo monitoramento automático, destacando o desempenho do sistema em identificar e resolver problemas antes que eles afetem a operação.
8. **Satisfação do cliente:** Avaliar a experiência dos passageiros por meio de questionários, feedback ou análise de reclamações, garantindo que as melhorias operacionais atendam às expectativas do público.

Atores

- **Persona Principal:**
 - **Nome:** Sérgio Ribeiro
 - **Idade:** 52 anos
 - **Profissão:** Gestor Operacional da CPTM, engenheiro de produção pela Escola Politécnica da USP
 - **Perfil:** Com 25 anos de experiência no setor ferroviário, Sérgio é conhecido por sua habilidade em otimizar operações e melhorar a eficiência na CPTM. Ele acredita que a análise de grandes volumes de dados é fundamental para melhorar a produtividade e a qualidade dos serviços, alinhando custos com a experiência do usuário.

- **Desafios:** Falta de ferramentas adequadas para analisar grandes volumes de dados, limitando a capacidade de identificar e prevenir falhas operacionais.
- **Objetivo:** Sérgio busca uma solução de Big Data que permita um monitoramento eficiente e preventivo das operações, ajudando a CPTM a se tornar uma referência em inovação tecnológica e qualidade de serviço.
- **Frase Motivadora:** "A eficiência operacional é fundamental, e a análise de dados pode nos ajudar a prever falhas antes que elas aconteçam."
- **Stakeholders:** Diretoria (stakeholder principal), membros do Conselho Administrativo, o Governo, a Secretaria de Transporte e a Gestão da CPTM (operação e manutenção).

Ações

1. **Análise de Relatórios Operacionais:** Sérgio pode utilizar os relatórios gerados pela solução para monitorar o desempenho das estações em tempo real. Isso permitirá que ele identifique rapidamente padrões de anomalias e tome decisões informadas sobre a operação, como ajustar horários de manutenção ou redirecionar trens.
2. **Implementação de Ações Corretivas:** Ao receber notificações de anomalias nas sincronizações das portas, Sérgio poderá acionar sua equipe para investigar a situação imediatamente. Ele poderá desenvolver um protocolo claro sobre como agir, garantindo que a resposta a problemas seja rápida e eficaz.
3. **Reuniões de Feedback e Melhoria Contínua:** Sérgio pode organizar reuniões regulares com sua equipe para discutir as descobertas a partir das análises e relatar como as ações corretivas impactaram a operação. Essas reuniões servirão para colher feedback e propor melhorias contínuas nos processos operacionais, garantindo que a solução se mantenha relevante e eficaz.
4. **Acompanhamento de Tendências de Fluxo:** Sérgio pode utilizar dados históricos de movimentação de passageiros para identificar mudanças no comportamento de uso das estações ao longo do tempo, como crescimento em horários não tradicionais ou redução em dias específicos. Com isso, ele pode propor ajustes estratégicos na operação e até reavaliar a necessidade de campanhas para aumentar o uso em horários de baixa demanda.
5. **Definição de Prioridades de Investimento:** Baseado nos relatórios sobre falhas frequentes por linha ou estação, Sérgio pode priorizar os recursos destinados à manutenção e atualização de equipamentos. Por exemplo, ele pode justificar a substituição de sensores com desempenho crítico ou sugerir melhorias nas estações com maior incidência de problemas.
6. **Criação de Indicadores Personalizados:** A partir da plataforma, Sérgio pode estabelecer novos KPIs que atendam a necessidades específicas, como o tempo médio de reparo após notificações de falhas ou a eficiência operacional de diferentes turnos. Esses indicadores podem ser incorporados nos relatórios para guiar novas metas de desempenho.

Valores

A implementação deste sistema visa gerar valor por meio de:

- **Redução de Custos Operacionais:** Diminuindo o número de falhas não detectadas ou corrigidas tarde, resultando em menos reparos caros ou interrupções de serviço.

- **Maior Segurança:** Detectando e corrigindo falhas operacionais antes que afetem a segurança dos passageiros e/ou da operação.
- **Eficiência Operacional:** Otimizando a movimentação de passageiros e a alocação de trens com base nos dados.

Riscos

- **Falhas na Coleta de Dados:** Se houver falhas ou problemas nos sensores, a qualidade dos dados pode ser comprometida.
- **Integração:** Pode haver dificuldades em integrar este sistema com os sistemas já existentes de monitoramento e operação da CPTM.
- **Dependência de Dados do Fabricante:** Como os dados da caixa preta são oriundos do fabricante do trem, pode haver limitações na flexibilidade e personalização da coleta de dados.

Performance/Impacto

- **Impacto no Passageiro:** Redução do tempo de espera durante embarques/desembarques, menor incidência de falhas operacionais, e uma operação geral mais eficiente e segura.
- **Eficiência Operacional:** Melhoria no agendamento de trens e na alocação de recursos, com base em dados mais precisos e completos sobre uso e performance dos trens.
- **Aumento de Receita:** Uma operação mais eficiente pode atrair mais passageiros, reduzir custos e aumentar a receita ao longo do tempo.

Em conclusão, o Data Product Canvas é uma ferramenta importante para projetos de Big Data, pois traz uma visão clara do produto a ser desenvolvido. Ele facilita a comunicação entre os stakeholders e a equipe, garantindo que as necessidades reais sejam atendidas e que todos estejam na mesma página à nível de compreensão do escopo, objetivo e propósito do projeto. Ao mapear ações estratégicas e focar na entrega de valor contínuo, o DPC minimiza o risco de que soluções inovadoras sejam subutilizadas, garantindo que as decisões sejam *data driven* e contribuam para a eficiência operacional contínua.

2. Arquitetura Macro

2.1. Componentes da Arquitetura

Nossa arquitetura é organizada em camadas, cada qual com uma função clara: reunir dados brutos, transformá-los, analisar informações e, finalmente, disponibilizá-las de forma simples e útil. Abaixo, estão descritos os principais componentes envolvidos e suas atribuições, mostrando o caminho completo que os dados percorrem até virarem insights.

- **Camada de Coleta (Bronze):** Aqui é onde tudo começa. Dados brutos. Muitas vezes desalinhados, incompletos ou em formatos distintos, eles chegam ao nosso repositório central (Data Lake). É o ponto

de entrada de informações vindas de diversas fontes, sejam elas bancos de dados internos, arquivos CSV, sistemas de monitoramento ou APIs.

- **Validação e Limpeza (Prata):** Após a coleta, entram em cena processos de ETL executados por ferramentas como AWS Glue ou AWS Lambda. Nesse estágio, asseguramos que as informações sejam consistentes e de qualidade. Dados duplicados são removidos, tipos de campos são padronizados, e validações com Pydantic garantem que nada fora do padrão chegue à etapa seguinte. Ao final, temos um conjunto de dados mais confiável e pronto para análises, ainda dentro do Data Lake.
- **Armazenamento e Modelagem (Ouro):** Agora que os dados foram refinados, eles são estruturados em um Data Warehouse otimizado para consultas analíticas. É nessa camada que entram tecnologias como o ClickHouse (ou outra solução OLAP), que aceleram e facilitam a realização de análises complexas, agregações e comparações históricas.
- **Processamento e Análise Distribuída:** Para extrair valor real dos dados, usamos o Apache Spark (rodando em AWS EMR) ou tecnologias equivalentes, capazes de processar grandes volumes de dados em paralelo. Isso significa analisar grandes quantidades de informações com rapidez, gerando estatísticas e indicadores que serão a base de insights valiosos.
- **Visualização e Integração (Ródio):** Por fim, todo esse trabalho de bastidor se materializa em dashboards e relatórios interativos. Ferramentas como o Streamlit entram em cena para dar ao usuário uma interface intuitiva, permitindo visualizar tendências, gargalos e oportunidades ocultas nos dados. É a "vitrine" do pipeline, onde gestores e analistas podem tomar decisões informadas e rápidas, sem precisar conhecer a fundo a infraestrutura por trás.

2.2 UML de Componentes

O diagrama de componentes UML descreve a organização e as interações entre os componentes de um sistema de software. O mesmo detalha como módulos, bibliotecas e outras partes do sistema se relacionam, destacando as dependências e interfaces possíveis para a comunicação. Para fins de UML 2.0, o termo "componente" refere-se a um módulo de classes que representa sistemas ou subsistemas independentes com capacidade de interagir com o restante do sistema.

Para isso, existe uma abordagem de desenvolvimento em torno de componentes: o desenvolvimento baseado em componentes (CBD). Nela, o diagrama de componentes identifica os diferentes componentes para que todo o sistema funcione corretamente. ([Lucidchart, 2024](#))

Em resumo, o UML de Componentes:

1. Imagina a estrutura física do sistema;
2. Presta atenção aos componentes do sistema e como eles se relacionam;
3. Enfatiza o comportamento do serviço quanto à interface.

O diagrama de componentes UML descreve a organização e as interações entre os componentes de um sistema de software. O mesmo detalha como módulos, bibliotecas e outras partes do sistema se relacionam, destacando as dependências e interfaces possíveis para a comunicação. Para fins de UML 2.0, o termo "componente" refere-se a um módulo de classes que representa sistemas ou subsistemas independentes com capacidade de interagir com o restante do sistema.

Para isso, existe uma abordagem de desenvolvimento em torno de componentes: o desenvolvimento baseado em componentes (CBD). Nela, o diagrama de componentes identifica os diferentes componentes para que todo o sistema funcione corretamente. ([Lucidchart, 2024](#))

Em resumo, o UML de Componentes:

1. Imagina a estrutura física do sistema;
2. Presta atenção aos componentes do sistema e como eles se relacionam;
3. Enfatiza o comportamento do serviço quanto à interface.

Arquitetura Medallion

A arquitetura medallion, ou medalhão, descreve uma série de camadas de dados que denotam a qualidade dos dados armazenados no Lakehouse.

Essa arquitetura garante a atomicidade, consistência, isolamento e durabilidade à medida que os dados passam por várias camadas de validações e transformações antes de serem armazenados em um layout otimizado para análise eficiente. Os termos bronze (bruto), prata (validado) e ouro (enriquecido) descrevem a qualidade dos dados em cada uma dessas camadas. ([Microsoft, 2024](#))

Figura 4 - Fluxo Geral do desenvolvimento da Solução



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

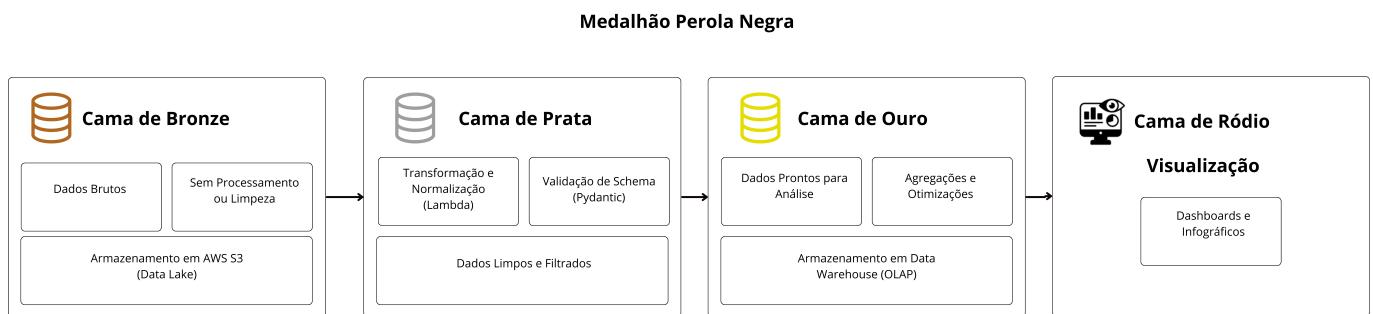
Para explicar como construímos nossa solução, acima temos uma imagem do Fluxo Geral de Desenvolvimento, sendo o mesmo composto pelas etapas:

1. **Camada de Dados**: Os dados são coletados de diversas tabelas, como Estacao, Trem_Passageiros, Intervalos_Dia e Mov_Periodo.
2. **Ingestão de Dados**: Os dados brutos obtidos das tabelas são armazenados em um repositório central.
3. **AWS S3 Data Lake**: Os dados são carregados em um data lake na AWS S3, onde passam por um processo de validação de esquema utilizando o Pydantic.
4. **Camada de Validação**: Os dados passam por processos de limpeza, deduplicação e conversão para formatos apropriados.
5. **Transformação com Lambda**: Os dados são carregados e transformados utilizando a função AWS Lambda para prepará-los para análises mais profundas.
6. **Processamento**: Os dados transformados são processados utilizando o Apache Spark, que possibilita processamento em larga escala.

7. **Data Warehouse OLAP:** Os dados processados são armazenados em um data warehouse projetado para consultas analíticas (OLAP).
 8. **Dashboards e Visualizações Streamlit:** Os resultados são apresentados por meio de dashboards interativos desenvolvidos com Streamlit.
-

Abaixo, passaremos por cada uma das camadas, bronze, prata, ouro e ródio, desenhadas para o grupo Pérola Negra considerando o foco do time nos dados de caixa preta do trem.

Figura 5 - Medalhão Perola Negra



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Camada de Bronze (Data Lake)

Armazena todos os dados de sistemas de origem externa. As estruturas da tabela desta camada correspondem às estruturas da tabela "*as is*" (como estão) no sistema de origem, juntamente com metadados de colunas adicionais, como data de carregamento, ID do processo etc.

Na Camada de Bronze do projeto "Medalhão Pérola Negra", foi trabalhado com dados fornecidos através do MinIO, uma solução de armazenamento compatível com o AWS S3 que atua como nosso Data Lake para dados brutos. Esse armazenamento inicial continha arquivos em diversos formatos, além do comum CSV. Essa diversidade de formatos exigiu um trabalho inicial de conversão e padronização.

Para garantir uma análise consistente dos dados na camada de transformação, foi necessário realizar uma etapa de conversão preliminar. Utilizando Jupyter Notebooks e ferramentas de manipulação de dados em Python, como Pandas, realizamos a extração e transformação dos dados para o formato CSV, facilitando o entendimento e permitindo uma análise exploratória mais fluida. Essa preparação inicial foi fundamental para identificar padrões, estruturar os dados e fazer as limpezas básicas necessárias antes de enviá-los para a Camada de Prata, onde são realizados os processos de transformação e validação.

Camada de Prata (Transformação e Normalização)

A Camada de Prata é onde ocorre a transformação e a limpeza dos dados. Nesta etapa, se inicia o processo de ETL (Extração, Transformação e Carga). Durante a transformação, os dados passam por normalizações e padronizações, o que inclui corrigir erros, remover duplicatas, converter tipos de dados e aplicar regras de negócio básicas. Após essa transformação, aplicamos o Pydantic, uma biblioteca Python que valida o schema dos dados, ou seja, garante que todos os registros estejam consistentes com o formato

esperado. Isso assegura que apenas dados limpos e formatados corretamente sigam para as próximas camadas.

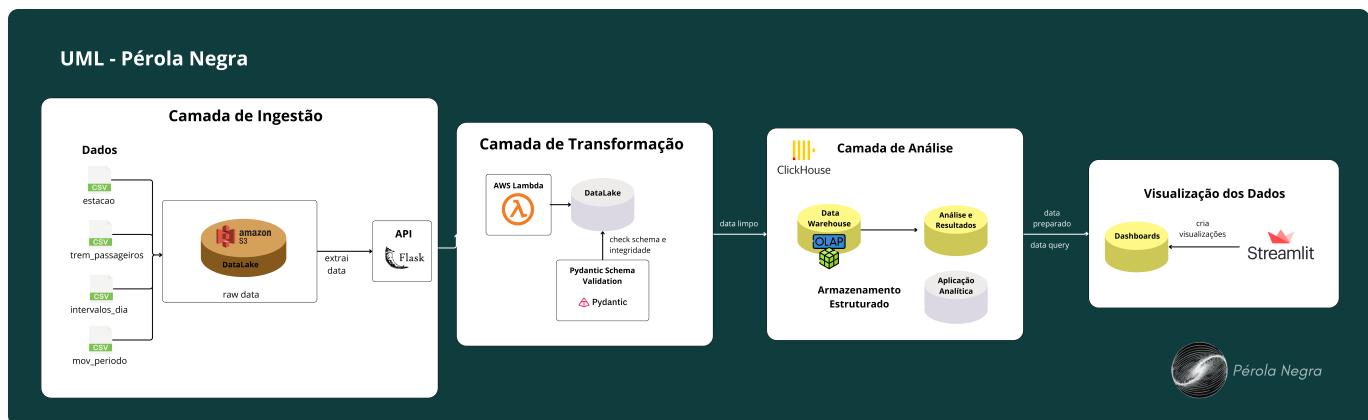
Camada de Ouro (Data Warehouse)

Na Camada de Ouro, os dados já transformados e validados na etapa anterior são agregados e otimizados para suportar análises mais avançadas e frequentes. Aqui, armazenamos esses dados em um Data Warehouse, no caso, o ClickHouse, que é um banco de dados analítico de alta performance. Na Camada de Ouro, os dados são preparados e organizados de forma a facilitar consultas rápidas e análises aprofundadas, sendo formatados para atender a consultas OLAP (Online Analytical Processing), que permitem operações complexas, como agregações e filtros eficientes.

Camada de Ródio (Visualização)

A Camada de Ródio foi criada para destacar a visualização dos dados, a etapa final e muitas vezes a mais valiosa para a tomada de decisões. Nessa camada, foi escolhido o Streamlit, hospedado em um servidor, para criar dashboards e infográficos interativos. Streamlit permite que visualizações dinâmicas sejam construídas de forma a transformar os dados processados em insights acionáveis. O grupo Pérola Negra usou o metal ródio para representar essa camada devido ao seu alto valor, simbolizando a importância e o impacto dos dados visualizados e analisados para o negócio.

Figura 6 - UML de Componentes - Pérola Negra



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ingestão

A solução proposta para o pipeline de Big Data da CPTM começa com a Camada de Ingestão no AWS S3, onde dados brutos são armazenados diretamente no Data Lake, criando a base para a coleta de dados de várias fontes, como Estação, Trem_Passageiros, Intervalos_Dia, Mov_Período e Tipo_Embalagem. Essa coleta é feita por meio de uma API desenvolvida em Flask, que realiza a extração dos dados, enviando-os para o S3. A API é responsável por garantir que esses dados sejam capturados de forma precisa e segura, mantendo-os centralizados para o próximo estágio.

Transformação

Na transição para a Camada de Transformação, a solução utiliza o AWS Glue ou, alternativamente, AWS Lambda para executar o processo **ETL** (Extração, Transformação e Carga). Esta camada transforma os dados brutos em um formato mais utilizável, aplicando deduplicação, limpeza, conversão de tipos e normalização dos dados. Além disso, para garantir a qualidade dos dados, o Pydantic é usado para validar o schema e assegurar que os dados estejam no formato correto antes de avançarem no pipeline. Testes de consistência e integridade são executados dentro do diretório /test, verificando a conformidade e estabilidade dos dados transformados. Esse processo resulta em dados limpos e preparados, que são armazenados na Camada Prata do Data Lake.

Análise

Na Camada de Análise, o AWS EMR (Elastic MapReduce) com suporte de Apache Spark e Hadoop processa os dados transformados para análises distribuídas e cálculos estatísticos descritivos. Aqui, os dados da camada Prata são convertidos em informações significativas e análises valiosas, sendo preparados para consumo em dashboards e relatórios.

Visualização

Para a Visualização dos Dados, a ferramenta Streamlit é hospedada em um servidor, permitindo criar dashboards e infográficos intuitivos e visualmente atrativos para os usuários. Para visualizações avançadas, uma ferramenta open-source é utilizada em um container dedicado, facilitando a visualização interativa e dinâmica dos dados prontos para análise.

O Armazenamento Estruturado é feito em um Data Warehouse OLAP, armazenando os dados prontos para consultas analíticas, permitindo que a Aplicação Analítica acesse os dados para fornecer uma interface de interação com relatórios e infográficos finais, otimizando a gestão e tomada de decisões na CPTM.

3. Processo de ETL

ETL é a sigla para o processo de extrair, transformar e carregar. É uma forma tradicionalmente aceita para que as organizações combinem dados de vários sistemas em um único banco de dados, repositório de dados, armazenamento de dados ou data lake. O ETL pode ser usado para armazenar dados legados, ou, o que é mais comum, agregar dados para analisar e impulsionar as decisões de negócios. ([Google Cloud, 2024](#))

Por meio do ETL, é possível definir a qualidade dos dados e a forma como eles são manipulados a fim de transformá-los em uma informação inteligível e confiável.

O processo é composto por três etapas distintas:

- **Extração:** Consiste em coletar dados de diversas fontes, como bancos de dados, APIs, ou arquivos externos. O objetivo é centralizar as informações necessárias para análise em um único lugar, garantindo que sejam obtidos dados de qualidade e que as fontes de dados sejam confiáveis.
- **Transformação:** Nessa etapa, os dados são limpos, organizados e transformados para que possam ser utilizados de forma consistente. As transformações podem incluir a remoção de duplicatas, tratamento

de valores ausentes, normalização e agregação dos dados, tudo para que estejam prontos e adequados para o propósito de análise.

- **Carregamento:** Após a transformação, os dados são carregados no destino final, geralmente um Data Warehouse (OLAP), onde ficarão disponíveis para consultas e análises. É importante garantir que o processo de carregamento seja eficiente e que a integridade dos dados seja preservada ao longo do processo.

ETL

Arquitetura e Fluxo do Pipeline de Dados

Nesta seção, será apresentado o fluxo geral e as funcionalidades principais do código responsável pelo pipeline de ingestão de dados. Esse pipeline foi projetado para extrair dados armazenados em arquivos .parquet de um bucket S3, transformá-los e inseri-los em uma base ClickHouse, organizando e monitorando o processo para garantir consistência e rastreabilidade.

Estrutura Organizacional do Pipeline

O pipeline inicia com a estruturação de uma pasta denominada schemas, onde são definidos modelos de dados específicos para diferentes tipos de informações que serão processadas. Esses modelos – TrensPassageirosModel, IntervalosDiaModel, EstacaoModel, MovPeriodoModel e TipoEmbarqueModel – descrevem a estrutura de cada conjunto de dados e as respectivas especificidades, garantindo que todos os dados atendam aos requisitos do sistema final.

Função `get_parquet_files`: Identificação de Arquivos de Dados

A função `get_parquet_files` atua na identificação dos arquivos .parquet armazenados no bucket perola-negra. Ela realiza uma busca para localizar todos os arquivos relevantes que serão importados para ClickHouse. Esse processo inicial assegura que o pipeline tenha uma lista completa dos dados que precisam ser processados.

Função `convert_to_unix`: Padronização Temporal

Para manter a consistência dos dados, a função `convert_to_unix` transforma qualquer dado temporal em um formato Unix, facilitando a manipulação e interpretação. Esta etapa é essencial para garantir que todos os dados compartilhem uma linguagem temporal comum ao serem inseridos no ClickHouse, minimizando problemas de compatibilidade e processamento.

Função `read_parquet_and_insert_to_clickhouse`: Execução da Ingestão de Dados

Esta função gerencia o processo de leitura, transformação e inserção dos dados. Suas principais etapas são:

- *Criação de Tabelas:* A função inicia criando uma tabela no ClickHouse chamada grupo5.data_ingestion para armazenar os dados importados, com base nas estruturas previamente definidas.
- *Leitura e Preparação de Dados:* Cada arquivo .parquet é lido e convertido para o formato compatível com ClickHouse.

- *Inserção Condicional*: A função valida e insere os dados de acordo com o tipo (TrensPassageiros, IntervalosDia, etc.), garantindo que cada conjunto seja identificado corretamente no ClickHouse.
- *Validação e Tratamento de Erros*: O pipeline inclui um sistema de validação que emite alertas para dados inválidos ou com estrutura inconsistente, mantendo a integridade e a confiabilidade do pipeline.
- *Registro de Logs*: Todos os eventos, erros e sucessos são documentados em um sistema de observabilidade (log_observability), permitindo a auditoria e o acompanhamento das operações.

Função `ingest_data`: Coordenação Geral da Ingestão

Como principal orquestradora do pipeline, a função `ingest_data` executa a coleta, transformando e transferindo os dados para o ClickHouse. Ela percorre cada bucket identificado, processa os arquivos .parquet e, por fim, chama a função `read_parquet_and_insert_to_clickhouse` para realizar a operação de ingestão completa, com relatórios no sistema de logs.

Além disso, a lógica do pipeline considera chaves primárias e partições ao criar e carregar dados nas tabelas do ClickHouse, garantindo que a distribuição dos dados seja equilibrada e a consulta seja otimizada. Por exemplo, no caso da tabela `grupo5.data_ingestion`, o esquema definido no modelo `TrensPassageirosModel` inclui uma chave primária com base em identificadores do trem e timestamps, permitindo buscas rápidas e filtragem eficiente. Assim, ao carregar os dados, o pipeline verifica a conformidade das colunas com o esquema, garante o tipo correto de cada campo (ex: inteiro, string, datetime) e assegura que informações temporais sejam padronizadas, evitando divergências entre diferentes fontes. Caso algum registro não atenda aos padrões, o pipeline registra o evento e pula para o próximo lote, mantendo assim a consistência dos dados.

Resumo do Pipeline

- **Pasta schemas**: Definição e estruturação dos modelos de dados.
- **get_parquet_files**: Identificação dos arquivos .parquet a serem processados.
- **convert_to_unix**: Padronização dos dados temporais em formato Unix.
- **read_parquet_and_insert_to_clickhouse**: Função de ingestão com validação, estruturação e registro.
- **ingest_data**: Controladora geral que realiza a coleta, processamento e documentação dos dados.

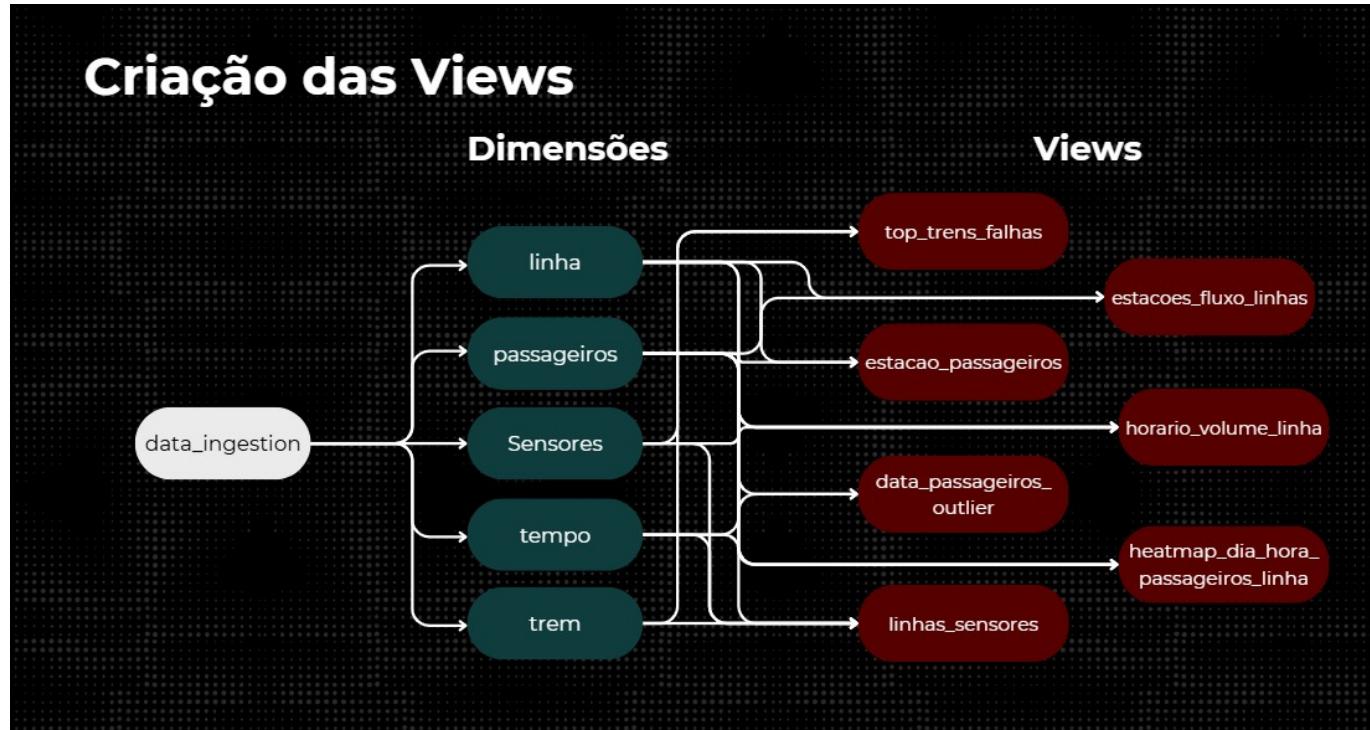
Este pipeline garante que cada conjunto de dados passe por uma estrutura organizada de extração, transformação e carga (ETL), com monitoramento e rastreabilidade, otimizando a integração com o ClickHouse e mantendo um histórico das operações realizadas.

3.1 Análise de Dados

Dimensões são estruturas fundamentais em um modelo de dados que organizam as informações em categorias ou grupos lógicos. No contexto deste projeto, as dimensões servem como bases estruturadas para agrupar, relacionar e acessar dados específicos de maneira eficiente. Elas permitem segmentar os dados em contextos relevantes para análises detalhadas e consultas direcionadas, facilitando a criação de relatórios e insights precisos.

As dimensões selecionadas para o projeto são fundamentais para a organização e análise dos dados no **DataApp**. Cada dimensão foi escolhida para garantir que as *views* criadas pudessem atender às necessidades operacionais e estratégicas da CPTM, melhorando a análise de dados relacionados ao transporte público. Abaixo é possível visualizar as dimensões e *views* utilizadas no projeto, seguidas de uma explicação detalhada.

Figura 7 - Dimensões e Views



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

1. **dimensao_estacoes:** Refere-se às estações da CPTM, incluindo detalhes como nome da estação, prefixos e atributos relevantes para análises. É essencial para mapear fluxos de passageiros e operações específicas por estação.
2. **dimensao_intervalo_dia:** Representa os intervalos diárias, como manhã, tarde e noite, sendo essencial para segmentar análises de operações ao longo do dia.
3. **dimensao_movimento_tempo:** Combina dados temporais com o movimento dos trens e dos passageiros. É crucial para análises de padrões operacionais e identificação de períodos críticos.
4. **dimensao_tipo_embarque:** Agrupa os tipos de embarque identificados, facilitando a análise do comportamento dos passageiros e das tendências de bilhetagem.
5. **dimensao_trens_passageiros:** Integra informações sobre os trens e os volumes de passageiros transportados, permitindo cruzamentos entre desempenho dos trens e ocupação.

As dimensões foram definidas levando em conta a lógica de chave estrangeira e chaves primárias para permitir junções eficientes entre tabelas. Por exemplo, a dimensão **dimensao_estacoes** possui um identificador único para cada estação (ex: `id_estacao`), que é utilizado nas *views* para relacionar eventos operacionais (como volume de passageiros ou falhas) a uma estação específica. Assim, ao consultar a *view* que mostra fluxo entre estações, o sistema realiza um join entre a tabela de fatos (com dados de fluxo) e a dimensão **dimensao_estacoes**, garantindo que os resultados sejam retornados rapidamente e de forma consistente. A normalização dessas dimensões segue padrões do tipo estrela (star schema), onde uma tabela fato central (como a de movimento de passageiros) se relaciona com dimensões que fornecem contexto. Esse design simplifica consultas OLAP, tornando a análise ágil e completa.

A criação de dimensões como tabelas facilita a manipulação e reutilização de dados estruturados, eliminando a necessidade de realizar extrações repetidas e otimizando o processamento de consultas.

As *views* criadas no **ClickHouse** organizam e sintetizam os dados coletados, sendo o alicerce para análises mais eficientes e direcionadas. Cada *view* foi desenvolvida para responder a questões operacionais específicas da CPTM, utilizando as dimensões previamente definidas.

1. **view_fluxo_entre_estacoes**: Relaciona o fluxo de passageiros entre diferentes estações. Baseia-se principalmente na dimensão **dimensao_estacoes** para entender os trajetos mais utilizados.
2. **view_heatmap_pessoas_por_linha**: Utiliza as dimensões **dimensao_estacoes** e **dimensao_movimento_tempo** para gerar mapas de calor com a distribuição de passageiros ao longo das linhas em períodos específicos.
3. **view_media_intervalo_operacao_por_dia**: Analisa a média dos intervalos de operação ao longo do dia, com base em dados temporais (**dimensao_intervalo_dia**).
4. **view_media_tempo_porta_aberta**: Apresenta o tempo médio em que as portas dos trens permanecem abertas, útil para otimizar a eficiência operacional. Relaciona-se com a dimensão **dimensao_movimento_tempo**.
5. **view_movimento_classificado_por_bilhete**: Classifica os movimentos dos passageiros com base nos tipos de bilhetes utilizados, utilizando a dimensão **dimensao_tipo_embarque**.
6. **view_sensores_por_data**: Relaciona dados coletados pelos sensores com a dimensão temporal, facilitando a análise de eventos ou falhas capturadas ao longo dos dias.
7. **view_tipos_bilhete_abundantes**: Identifica os tipos de bilhetes mais utilizados com base nos registros da dimensão **dimensao_tipo_embarque**.
8. **view_tipos_bilhete_por_dia**: Analisa o uso de bilhetes ao longo dos dias, cruzando dados temporais com os tipos de embarque.
9. **view_tipos_bilhete_por_semana**: Detalha o uso dos bilhetes durante a semana, permitindo identificar tendências sazonais e padrões de uso.

Cada *view* foi construída com base em consultas SQL otimizadas, que utilizam índices e partições presentes no ClickHouse. Por exemplo, em **view_fluxo_entre_estacoes**, a consulta utiliza filtragem por intervalos de tempo e junção com **dimensao_estacoes** para reduzir o conjunto de dados analisado, aumentando a velocidade da resposta. Com isso, a lógica interna da *view* garante que apenas as colunas necessárias sejam retornadas, diminuindo a carga no banco e melhorando a experiência do usuário final.

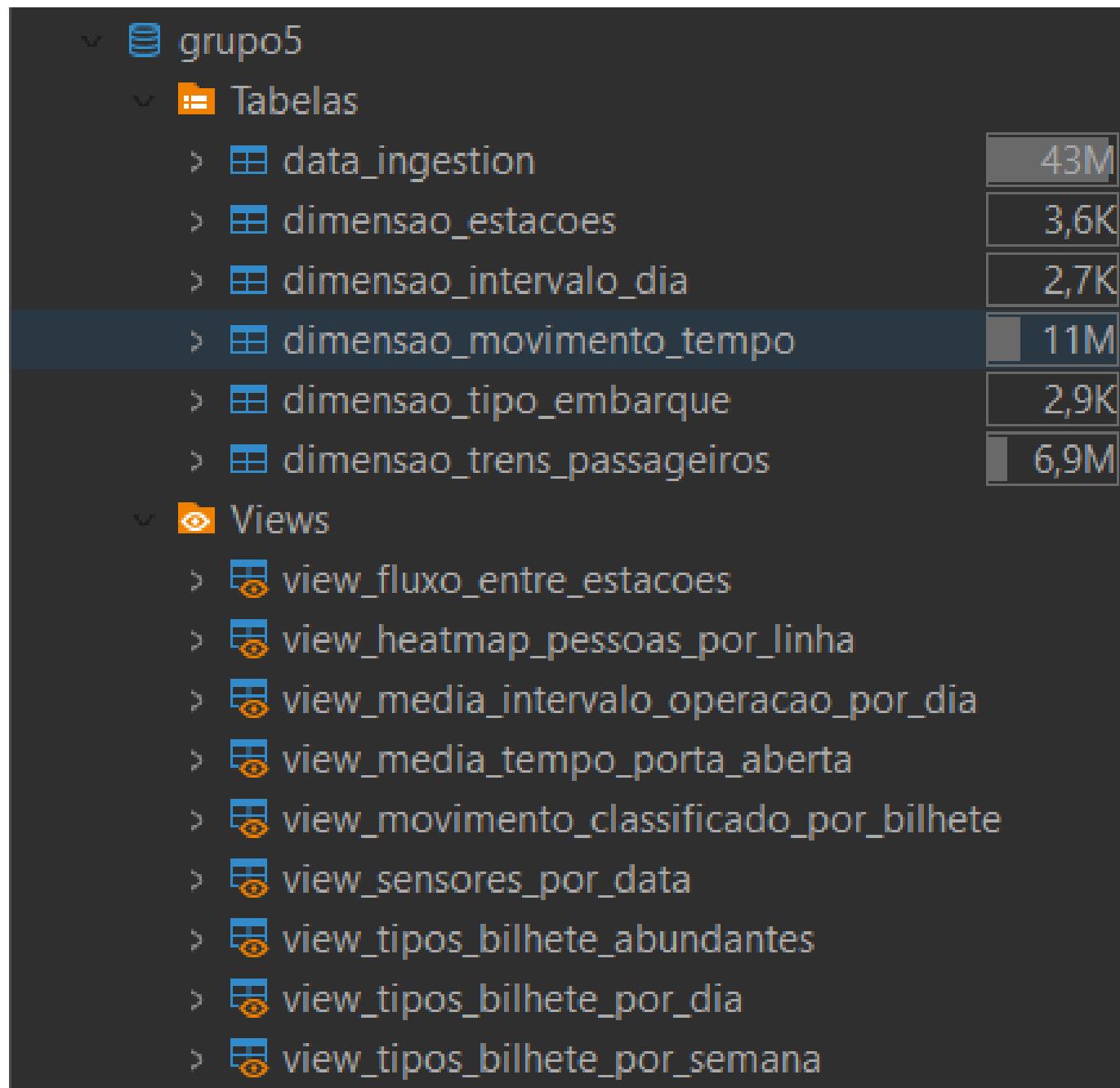
3.1.1 Criação das Planilhas

As tabelas de dimensões e *views* criadas no banco de dados **grupo5** foram desenvolvidas para organizar e estruturar os dados, permitindo análises alinhadas às necessidades operacionais. As dimensões mencionadas abordam aspectos essenciais do sistema, como:

- Estações e seus atributos (**dimensao_estacoes**)
- Intervalos temporais ao longo do dia (**dimensao_intervalo_dia**)
- Comportamento dos movimentos temporais (**dimensao_movimento_tempo**)
- Tipos de embarque identificados (**dimensao_tipo_embarque**)
- Dados específicos de trens e passageiros (**dimensao_trens_passageiros**)

Essas dimensões servem de base para cruzamentos e análises de dados mais específicas, permitindo insights para a operação da CPTM. Veja abaixo como está essa organização de forma mais visual.

Figura 8 - Dimensões e Views no DBeaver



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A Figura acima apresenta a estrutura geral do banco de dados, com destaque para as dimensões e *views* que suportam as análises realizadas no projeto. Por exemplo, a tabela *dimensao_estacoes* armazena informações sobre as estações e as linhas associadas, incluindo o nome da estação, prefixos e descrição das linhas.

Figura 9 - Exemplo de dimensao_linha

The screenshot shows the ClickHouse Management Interface. On the left, the schema browser displays various tables and views, including 'dimensao_estacoes' (3,6K), 'dimensao_intervalo_dia' (2,7K), 'dimensao_movimento_tempo' (11M), 'dimensao_tipo_embarque' (2,9K), 'dimensao_trens_passageiros' (6,9M), and several views under the 'Views' category. On the right, a query results table for 'view_fluxo_entre_estacoes' is shown, with columns 'id_estacao', 'tx_prefixo', 'tx_nome', and 'cd_estacao_bu|cluster'. The data includes rows for various stations like ÁGUA BRANCA, ANTONIO GIANETTI NETO, etc., across different clusters.

	id_estacao	tx_prefixo	tx_nome	cd_estacao_bu cluster
1	ABR	ÁGUA BRANCA		511 2
1	ABR	ÁGUA BRANCA		511 2
4	AGN	ANTONIO GIANETTI NETO		709 0
4	AGN	ANTONIO GIANETTI NETO		709 0
5	AJO	ANTONIO JOÃO		558 2
5	AJO	ANTONIO JOÃO		558 2
6	ARC	ARACARÉ		753 0
6	ARC	ARACARÉ		753 0
8	BFI	BALTAZAR FIDELIS		502 2
8	BFI	BALTAZAR FIDELIS		502 2
9	BFU	PALMEIRAS-BARRA FUNDA		517 2
9	BFU	PALMEIRAS-BARRA FUNDA		517 2
10	BRU	BARUERI		557 2
10	BRU	BARUERI		557 2
11	BRR	BERRINI		606 2
11	BRR	BERRINI		606 2
12	BTJ	BOTUJURU		512 2
12	BTJ	BOTUJURU		512 2
13	BAS	BRÂS		764 0

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Além das dimensões, as *views* são elementos importantes na organização dos dados para análise. A *view view_fluxo_entre_estacoes*, por exemplo, destaca os fluxos com base em registros dos sensores. Ela permite identificar os problemas mais frequentes e priorizar ações de manutenção bom base no volume.

Figura 10 - Exemplo de View top_trens_falhas

The screenshot shows the ClickHouse Management Interface. On the left, the schema browser displays various tables and views, including 'dimensao_estacoes' (3,6K), 'dimensao_intervalo_dia' (2,7K), 'dimensao_movimento_tempo' (11M), 'dimensao_tipo_embarque' (2,9K), 'dimensao_trens_passageiros' (6,9M), and several views under the 'Views' category. On the right, a query results table for 'view_fluxo_entre_estacoes' is shown, with columns 'estacao_inicio|estacao_fim|total_entradas|total_saídas'. The data includes rows for various station pairs with their total entry and exit counts.

	estacao_inicio estacao_fim	total_entradas	total_saídas
4	6	147543	224737
6	4	136324	122326
2	6	43673	23136
6	2	39319	74775
1	6	27712	11367
6	1	21464	41808

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Essas tabelas e *views* formam o alicerce do sistema de análise, fornecendo os dados que suportam a tomada de decisão operacional e estratégica que será exposto em infográficos. As dimensões fornecem a base estruturada, enquanto as *views* sintetizam os dados para análises específicas, como fluxo de passageiros, falhas de trens e desempenho operacional de linhas.

3.2 Cubo de Dados

O Prefect é uma plataforma de orquestração de workflows que permite a automação, monitoramento e gestão de tarefas complexas. Ele é amplamente utilizado para ETL (Extract, Transform, Load) e pipelines de dados, garantindo eficiência e controle em processos de atualização e análise. A principal vantagem do Prefect é sua capacidade de integrar diferentes fontes de dados e ferramentas, além de fornecer uma interface intuitiva para monitorar e gerenciar fluxos em tempo real.

No projeto da CPTM, o Prefect foi configurado para gerenciar a criação e atualização de todas as *views* no banco de dados ClickHouse. Essa configuração assegura que os dados sejam processados e organizados

automaticamente, eliminando a necessidade de intervenções manuais e garantindo que as informações estejam sempre atualizadas.

Cada view criada no ClickHouse está vinculada a um flow do Prefect, que encapsula as tarefas necessárias para sua criação ou atualização. Essas tarefas são configuradas para rodar em intervalos específicos ou sob demanda, dependendo da necessidade operacional.

A criação das views foi desenvolvida garantindo eficiência e clareza na organização das informações. Cada view é construída a partir de consultas SQL específicas, projetadas para transformar e estruturar os dados extraídos das tabelas de origem de forma consistente e otimizada. Como exemplo, abaixo está o código da view "top_trens_falhas".

```
from prefect import task
from config.connections import get_clickhouse_client
import os

@task(name="Create View Top Trens Falhas")
def create_top_trens_falhas_view():
    client = get_clickhouse_client()
    sql_query = """
        CREATE OR REPLACE VIEW grupo5.top_trens_falhas AS
        SELECT
            dt.train_id AS id_trem,
            ds.id_sensor AS id_sensor,
            ds.sensor_sts AS status_falha,
            COUNT(*) AS ocorrencias -- Conta as repetições
        FROM
            grupo5.dimensao_sensores AS ds
        JOIN
            grupo5.dimensao_trem AS dt ON ds.door_id = dt.door_id
        WHERE
            ds.sensor_sts != 0 -- Apenas falhas
        GROUP BY
            dt.train_id, ds.id_sensor, ds.sensor_sts
        ORDER BY
            ocorrencias DESC;
    """
    client.execute(sql_query)
    return "View 'top_trens_falhas' criada com sucesso!"
```

O código acima cria a view "top_trens_falhas" no ClickHouse usando Prefect para orquestração. A task `create_top_trens_falhas_view` executa uma consulta SQL que relaciona falhas dos sensores com os trens, filtrando apenas sensores com falhas, contando as ocorrências e organizando os resultados em ordem decrescente. A view é criada ou atualizada automaticamente, permitindo identificar os trens com maior número de falhas de forma eficiente.

Outras views seguem estruturas semelhantes, sendo possível de serem analisadas ao acessar o DBeaver de forma mais simples, como pode-se ver na imagem da “top_trens_falhas” a seguir.

Figura 11 - DBeaver - top_trens_falhas

The screenshot shows the DBeaver 24.2.5 interface with the title 'DBeaver 24.2.5 - top_trens_falhas'. The left sidebar shows a tree view of the database structure under 'Navegador de banco de dados'. The main area displays a table titled 'top_trens_falhas' with the following columns: Grade, id_trem, id_sensor, status_falha, and ocorrencias. The data is sorted by Grade. The table contains 24 rows of data, with the first few rows being:

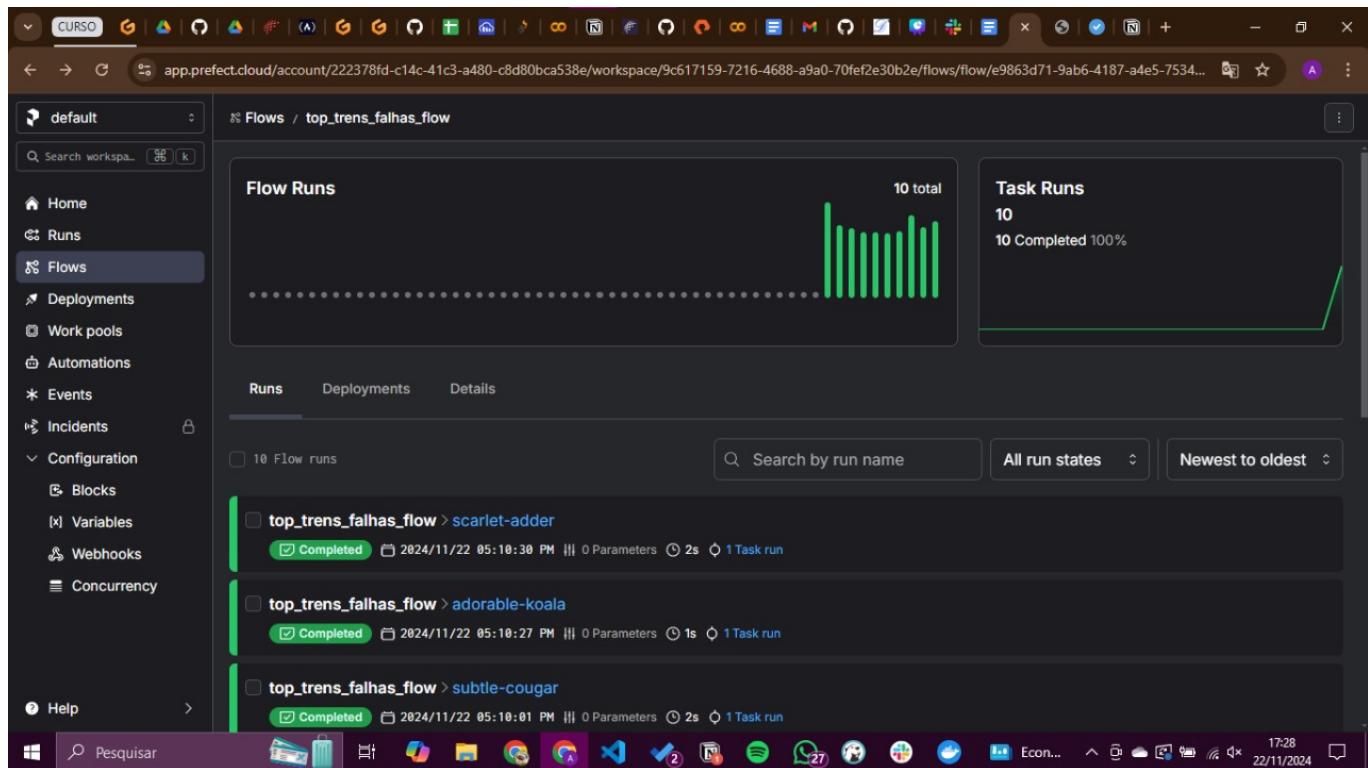
Grade	id_trem	id_sensor	status_falha	ocorrencias
1	-1	8	3	1.006.382
2	-1	16	3	928.968
3	-1	56	3	928.968
4	-1	48	3	851.554
5	-1	24	3	851.554
6	-1	4	3	851.554
7	-1	32	3	851.554
8	-1	64	3	851.554
9	-1	40	3	851.554
10	-1	80	3	774.140
11	-1	88	3	774.140
12	-1	96	3	774.140
13	-1	104	3	774.140
14	-1	72	3	774.140
15	-1	112	3	696.726
16	-1	120	3	696.726
17	-1	128	3	696.726
18	-1	12	3	696.726
19	-1	144	3	619.312
20	-1	176	3	619.312
21	-1	20	3	619.312
22	-1	208	3	619.312
23	-1	184	3	619.312
24	-1	152	3	619.312

At the bottom of the interface, there is a status bar showing 'Views: top_trens_falhas' and a system tray with icons for network, battery, and date/time.

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

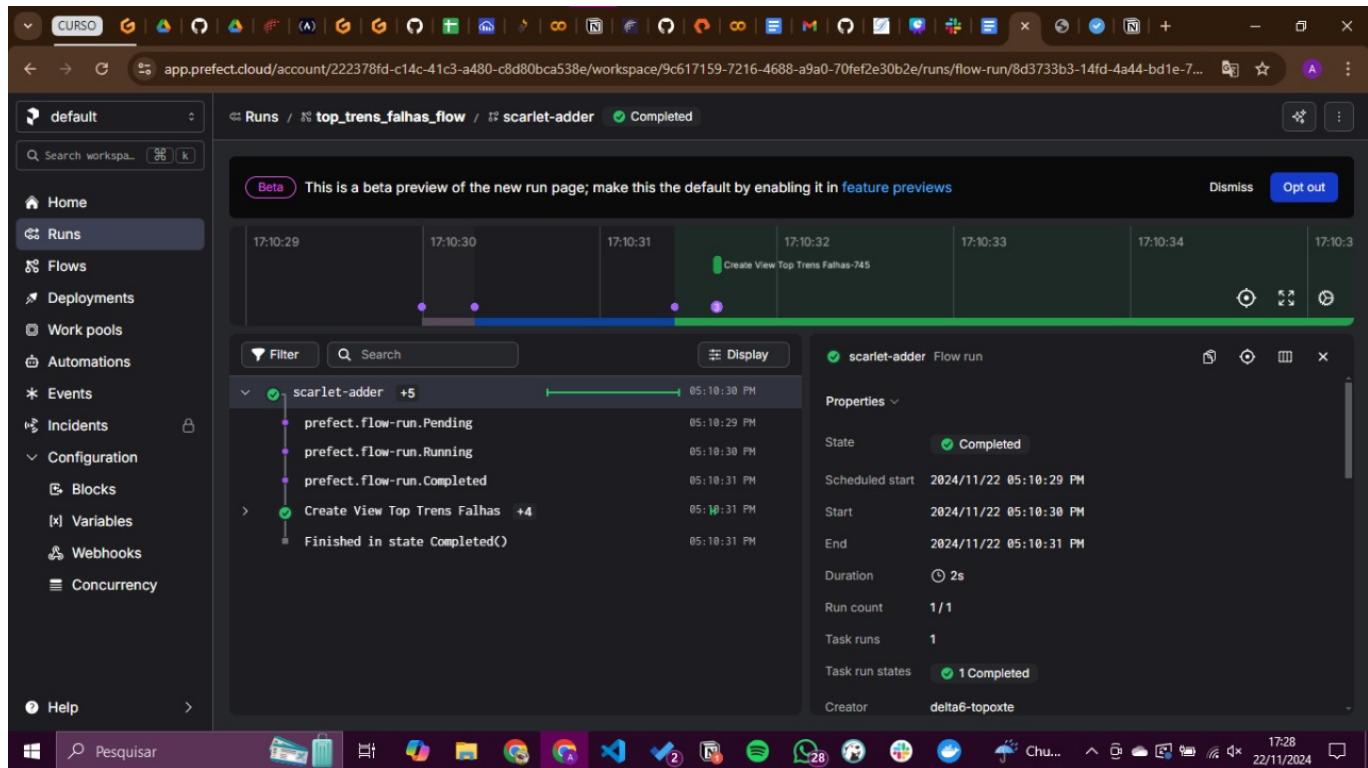
O Prefect oferece uma interface web que permite o monitoramento completo de todas as views configuradas no projeto. A plataforma registra logs detalhados de cada execução, incluindo informações sobre o status das tarefas, como erros ocorridos ou sucessos alcançados. Também é possível visualizar o estado atual dos flows, identificando quais já foram concluídos, quais estão em execução e quais aguardam na fila para serem processados. Em casos de falha, o Prefect facilita a reexecução manual ou automática do fluxo afetado, garantindo a continuidade do processo. Abaixo estão dois exemplos de visualização das views pelo Prefect.

Figura 12 - Histórico de Execuções Prefect



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 13 - Monitoramento de Execução Prefect



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

A primeira imagem exibe a visão geral de todas as execuções do flow `top_trens_falhas_flow`. No painel, é possível ver o gráfico de barras que representa o histórico de execuções, destacando o número total (10) e a taxa de conclusão (100%). Abaixo, há uma lista com as execuções detalhadas, mostrando informações como o

nome atribuído automaticamente a cada execução (e.g., scarlet-adder, adorable-koala), o estado final (Completed), a duração (em segundos), e o número de tarefas executadas em cada flow. Essa visualização permite que a equipe acompanhe a consistência e o sucesso das atualizações da view ao longo do tempo.

Partindo para a segunda imagem, ela apresenta a interface de execução de um flow específico no Prefect, chamado top_trens_falhas_flow. Nela, é possível visualizar a linha do tempo da execução, indicando estados como Pending, Running e Completed, cada um com um marcador temporal. À direita, estão as propriedades do flow, que incluem o estado final (Completed), o horário agendado para início, o horário de término, e a duração total da execução (2 segundos). Essas informações são fundamentais para monitorar a eficiência e o desempenho das execuções. A lista de tarefas do flow, como Create View Top Trens Falhas, também é exibida, destacando que todas foram concluídas com sucesso.

Essa capacidade de monitoramento garante que a equipe possa identificar e resolver problemas rapidamente, minimizando impactos operacionais. Além disso, a automação reduz o risco de inconsistências nos dados, uma vez que todas as etapas do pipeline são rastreáveis e repetíveis.

4. Análise de Impacto Ético

Nessa seção é descrito os possíveis impactos de big data na sociedade e no meio ambiente.

Introdução

Essa seção visa oferecer uma análise de impacto ético envolvendo cinco dimensões, que são:

1. Privacidade e proteção de dados;
2. Equidade e justiça;
3. Transparência e consentimento informado;
4. Responsabilidade social;
5. Viés e discriminação.

Nesse contexto, é explorado essas cinco dimensões trazendo uma análise crítica de como o projeto pode impactar a sociedade para a CPTM. Diante disso, as palavras de Desmond Tutu, arcebispo sul-africano e ativista contra o apartheid, vêm para reforçar essa responsabilidade: "se você é neutro em situações de injustiça, você escolheu o lado do opressor."([Globo](#), 2024) Por isso, torna-se extremamente importante uma análise crítica, com recortes voltados para grupos minoritários na sociedade, para que realmente seja possível exercer um impacto positivo.

Impactos em Meio Ambiente e Sociedade

A Companhia Paulista de Trens Metropolitanos (CPTM) desempenha um papel social na mobilidade urbana no estado de São Paulo, atendendo diariamente a cerca de 1,6 milhão de passageiros ([CPTM](#), 2023). Nos aspectos ambientais, a CPTM, que declara utilizar energia limpa, sendo assim com o DataApp de Big Data, capaz de reduzir ainda mais as emissões e o consumo de recursos por meio de análises proporcionada pelos 5 grupos da turma de 2023 de Sistemas de Informação, otimizando o uso de energia no mínimo desperdício ([CPTM](#), 2023). A gestão de resíduos será também aprimorada, favorecendo a reciclagem e reutilização. Contudo, o impacto do armazenamento excessivo de dados requer atenção, pois dados

irrelevantes em data centers elevam o consumo de energia e, consequentemente, as emissões de carbono (Caetano, 2020). A eliminação de dados não é recomendada, uma vez que todos os dados operacionais e ambientais têm potencial valor para a CPTM. Contudo, é importante considerar o ciclo de utilidade dos dados, que pode durar aproximadamente três anos. Após esse período, dados que não apresentem mais relevância para operações atuais podem ser movidos para um arquivo de dados inativos, onde não haverá consumo significativo de processamento. Para otimizar o processamento, é essencial que a equipe da CPTM esteja capacitada para gerenciar dados de forma eficiente, evitando sobrecargas desnecessárias no sistema, uma vez que o processamento pode se tornar o maior problema, tanto por capacidade e por custo.

Em termos sociais, a inclusão e a acessibilidade serão beneficiadas por uma mobilidade urbana mais eficiente, com a redução do tempo de espera e maior frequência dos trens, promovendo a inclusão social (CPTM, 2022). A análise demográfica permitirá compreender as necessidades dos usuários, assegurando que os serviços atendam melhor as populações vulneráveis. Além disso, o uso de análise de Big Data permite identificar com mais clareza os locais e condições onde ocorrem incidentes, facilitando a implementação de medidas para mitigá-los e aumentando a segurança dos passageiros (CPTM, 2023). A transparência sobre operações e segurança pode fortalecer a confiança do público na CPTM.

Para garantir que o projeto de Big Data da CPTM esteja em sintonia com sua política ambiental e práticas ESG, é essencial adotar uma abordagem holística e integrada. Essa abordagem vai além da conformidade com legislações ambientais e inclui a consideração dos impactos sociais e ambientais gerados em toda a cadeia de operações da CPTM. A política de sustentabilidade da empresa já abrange ações em prol da redução de emissões, otimização de recursos naturais e iniciativas voltadas para o bem-estar das comunidades atendidas (CPTM, 2023). O uso responsável de Big Data nesse contexto pode fortalecer essa estratégia, permitindo uma análise mais detalhada e preditiva dos impactos ambientais, eficiência energética e melhorias operacionais. Dessa forma, o projeto se torna não apenas uma ferramenta de análise, mas um meio de suportar a missão ESG da CPTM, potencializando o compromisso da empresa com uma visão de futuro mais sustentável e integrada (CPTM, 2023).

A implantação de uma solução de Big Data na CPTM oferece uma oportunidade estratégica para o aprimoramento tanto ambiental quanto social. Ao focar em áreas-chave como eficiência energética, gestão consciente de dados e inclusão social, a CPTM poderá não apenas otimizar suas operações internas, mas também desempenhar um papel ativo na construção de um futuro sustentável. Para garantir o sucesso e a sustentabilidade desse processo, será fundamental estabelecer metas claras e viáveis, além de documentar e monitorar continuamente os resultados. Esse acompanhamento constante permitirá não apenas assegurar o cumprimento dos objetivos, mas também antecipar e mitigar possíveis impactos negativos. Para maior profundidade sobre a conformidade do projeto com o impacto ético, veja a análise das 5 dimensões a seguir: Privacidade e proteção de dados, Equidade e justiça, Transparência e consentimento informado, Responsabilidade social, Viés e discriminação.

4.1. Privacidade e Proteção de Dados

Essa seção documenta os métodos de coleta, armazenamento e utilização de dados no projeto, garantindo que as práticas adotadas estejam em conformidade com as regulamentações de proteção de dados e respeitem os princípios de privacidade.

Coleta de Dados

- **Métodos de Coleta:**

- Os dados são capturados através de **sensores** distribuídos nas estações e trens, além de **bases de dados externas** que contribuem com informações para a melhoria operacional. A coleta é justificada pela necessidade de melhorar a experiência do usuário, otimizar operações e aumentar a eficiência geral do sistema de transporte.
- Caso não sejam coletados dados que possam identificar indivíduos, deve-se registrar que a coleta é de **dados anônimos** ou agregados, que são caracterizados como **dados não pessoais**. Exemplos incluem dados como padrões de uso e fluxo geral de passageiros.

- **Tipos de Dados Coletados:**

- **Dados pessoais:** informações que poderiam identificar um indivíduo diretamente, como nome e CPF, só serão coletados se absolutamente necessários e mediante conformidade com a legislação.
- **Dados não pessoais:** incluem informações agregadas e não identificáveis, como o número de passageiros em intervalos específicos, padrões de utilização e informações sobre fluxo de transporte.
- **Transparência no processo de coleta:** todos os usuários serão informados sobre o uso de seus dados por meio de notificações claras e acessíveis. Isso inclui detalhes sobre a finalidade da coleta, o período de retenção e as medidas de segurança adotadas.

Armazenamento de Dados

- **Procedimentos de Armazenamento:**

- Os dados são armazenados na AWS S3, garantindo escalabilidade e alta disponibilidade para grandes volumes. Para garantir a flexibilidade, dados estruturados são salvos em CSV, enquanto dados semi-estruturados utilizam JSON, que facilita integrações futuras e análises avançadas. A arquitetura implementa o ClickHouse para consultas analíticas rápidas e eficientes, otimizando o processamento de grandes volumes de dados em tempo real. A ferramenta DBeaver está integrada ao ClickHouse, possibilitando a conexão direta para consulta e visualização dos dados, facilitando a exploração e análise de informações de maneira intuitiva. Para mais detalhes sobre a arquitetura e uso dos componentes, consulte a seção 2.2 UML de componentes.

- **Medidas de Segurança:**

- A segurança dos dados é garantida pelos recursos oferecidos pela **AWS**, que implementa proteção de dados de ponta para armazenamento, além de políticas rígidas de gerenciamento de identidades e permissões de acesso.
- **Controles de acesso rigorosos** são adotados, com restrições configuradas para que apenas pessoal autorizado tenha acesso aos dados sensíveis, utilizando autenticação segura e senhas fortes.
- Auditorias regulares e práticas de monitoramento contínuo são realizadas, assegurando a identificação rápida de tentativas de acesso não autorizado e promovendo a integridade dos dados.

Uso de Dados

- **Finalidades do Uso:**

- Os dados são empregados para fins específicos, como **melhoria nas operações de transporte**, otimização de horários, manutenção preventiva, e análise de desempenho dos serviços.
- Garantimos que os dados serão usados estritamente para os fins informados aos usuários, com relatórios regulares que asseguram a transparência no uso e divulgação.

- **Compartilhamento com Terceiros:**

- Caso ocorra compartilhamento de dados com terceiros, este será limitado a **parceiros estratégicos**, assegurando que estejam implementadas condições rigorosas para a proteção dos dados, sempre em conformidade com a legislação vigente.

Conformidade com a LGPD

- **Práticas de Conformidade:**

- Quando aplicável, será obtido o **consentimento explícito** dos usuários para a coleta e o uso de quaisquer dados pessoais.
- Foi implementado a política de **minimização de dados**, que assegura que somente os dados necessários ao projeto sejam coletados e mantidos.

- **Práticas de Proteção e Segurança:**

- São recomendadas auditorias e monitoramentos regulares para evitar acessos não autorizados, e todos os funcionários devem receber **treinamentos periódicos** sobre proteção de dados e práticas de segurança.
- Relatórios de transparência sobre o uso dos dados e medidas de segurança devem ser emitidos periodicamente por conta de CPTM, visando a conformidade com a LGPD e promovendo a confiança dos usuários.

A documentação deste processo deverá ser revista periodicamente, garantindo que todas as práticas se mantenham atualizadas em relação às regulamentações vigentes e às melhores práticas do setor, conforme indicado pela política de sustentabilidade da CPTM ([CPTM](#), 2023).

4.2. Equidade e Justiça

Objetivo: A equidade e a justiça social são princípios fundamentais que devem guiar qualquer iniciativa em uma empresa pública, especialmente em uma organização como a Companhia Paulista de Trens Metropolitanos (CPTM), cuja missão é fornecer serviços essenciais para a população. No contexto de um projeto de centralização de dados em um sistema de big data, é indispensável assegurar que o tratamento, análise e aplicação desses dados estejam alinhados com esses princípios. Isso inclui garantir que as informações coletadas correspondam à realidade e que não sejam perpetuados vieses que possam levar a decisões injustas ou à marginalização de determinados grupos.

A inclusão de dados que retratem de forma fiel as características e necessidades de diferentes segmentos da sociedade, como as pessoas com deficiência (PCD), é essencial para reduzir desigualdades. Dessa forma, significa estruturar o pipeline de dados para evitar preconceitos já existentes, como racismo e outras formas de discriminação, garantindo que os benefícios do projeto sejam realmente acessíveis a todos. Esse cuidado é indispensável para que o transporte público da CPTM seja mais justo, inclusivo e alinhado com os princípios de equidade que a empresa representa.

- **Identificação de Impactos:**

- Passageiros: Incluem-se aqui usuários regulares e específicos, como idosos, PCDs e pessoas em situação de vulnerabilidade social. O projeto pode afetar diretamente a experiência desses grupos ao influenciar decisões sobre infraestrutura, alocação de recursos e otimização de serviços.
- Equipes Operacionais: Com a centralização dos dados, mudanças nos processos de trabalho podem ocorrer, afetando a dinâmica e a capacitação de funcionários.
- Gestores e Planejadores: A análise centralizada permitirá decisões mais embasadas, mas pode também amplificar vieses se os dados não forem devidamente tratados.
- Barreiras de Acesso: Se os dados utilizados para o planejamento não representarem adequadamente a diversidade dos usuários, algumas populações, como as que residem em áreas periféricas ou possuem necessidades específicas, podem ser prejudicadas.

- **Estratégias de Mitigação:**

- Inclusão de Dados Diversos: Garantir que o banco de dados inclua informações detalhadas sobre todos os perfis de usuários, incluindo PCDs, idosos e grupos socioecononomicamente vulneráveis. A adição de dados fornecidos pela CPTM, especificamente relacionados a PCDs, é um exemplo positivo de como garantir a representatividade.
- Indicadores de Equidade: O grupo de passageiros da CPTM está liderando os esforços para monitorar e avaliar os indicadores diretamente relacionados à equidade. Isso inclui métricas como acessibilidade, frequência de trens e qualidade do atendimento em áreas com maior vulnerabilidade social.
- Auditorias Éticas e Técnicas: Realizar revisões periódicas no pipeline de dados para identificar e corrigir vieses que possam surgir na coleta, armazenamento e análise das informações.

4.3. Transparência e Consentimento Informado

Para assegurar que todas as partes interessadas no projeto de Big Data da CPTM tenham acesso claro e transparente às informações sobre o uso dos dados, garantindo que o consentimento seja obtido de forma informada e voluntária.

- **Comunicação com Usuários:**

- Fica sob responsabilidade da CPTM informar aos usuários e stakeholders da mesma sobre a coleta e o uso de dados por meio de **comunicados visuais, campanhas de conscientização e avisos em plataformas digitais** utilizadas pela CPTM. As informações são apresentadas de maneira clara e acessível, utilizando exemplos práticos, como cartazes explicativos nas estações, notificações nos aplicativos oficiais da CPTM e vídeos educativos em monitores dentro dos trens.
- Para garantir a atualização constante, também fica sobre responsabilidade da CPTM, as políticas de privacidade e os guias informativos, que devem ser revisados semestralmente e disponibilizados tanto em meios digitais quanto físicos, para fácil acesso por todas as partes envolvidas.
- Para reforçar o entendimento, as políticas de privacidade são traduzidas para linguagem simples e incluem exemplos cotidianos do impacto do uso de dados na melhoria dos serviços, como otimização de horários e manutenção preventiva.

- **Consentimento:**

- O consentimento é formalizado através de um **Termo de Consentimento Informado**, que especifica de maneira simplificada e visualmente acessível, detalhando quais dados serão coletados, suas finalidades e os direitos dos cidadãos sobre o uso e proteção de seus dados. Esse termo é elaborado em conformidade com a Lei Geral de Proteção de Dados (LGPD) e normas estaduais, reforçando o compromisso com a privacidade e os direitos dos cidadãos. Um exemplo desse documento pode ser encontrado na seção de Anexos, como o Anexo I.
- Todos os registros de consentimento são armazenados em um sistema seguro e auditável por responsabilidade da CPTM, garantindo que possam ser revisados futuramente para fins de conformidade e segurança. O processo de revogação de consentimento pode ser realizado diretamente por meio do aplicativo oficial da CPTM ou nas bilheterias físicas, garantindo que todos os cidadãos, independentemente de familiaridade com tecnologia, consigam gerenciar suas preferências de privacidade.
- Os dados coletados, sempre que possível, são apresentados de forma agregada e anonimizada, minimizando riscos de exposição indevida e priorizando a proteção dos indivíduos. Além disso, os cidadãos são informados previamente sobre qualquer mudança significativa nas práticas de coleta ou uso de dados por meio de e-mails, mensagens no aplicativo oficial ou comunicados em estações.
- A CPTM disponibiliza um canal de atendimento especializado para esclarecimentos sobre o consentimento e uso de dados. Esse serviço é oferecido por meio de plataformas de fácil acesso para que qualquer cidadão possa obter informações claras e tirar dúvidas, promovendo transparência e confiança no uso de dados em um serviço público ([CPTM. \(s.d.\)](#)).

Como afirmou o líder espiritual **Dalai Lama**, "A falta de transparência resulta em desconfiança e um profundo sentimento de insegurança" ([Lama, 2024](#)). Esse pensamento reforça a importância de práticas transparentes na coleta e uso de dados, que promovem a confiança entre a CPTM e seus stakeholders, essenciais para o sucesso do projeto.

A implementação dessas práticas não apenas a conformidade com a legislação aplicável, mas também promove uma comunicação efetiva e acessível, criando uma cultura de respeito à privacidade e à autonomia dos usuários, que são fundamentais em projetos de dados na era digital.

4.4. Responsabilidade Social

A ética dos dados refere-se aos princípios que orientam o uso responsável e justo das informações, assegurando que sua coleta, armazenamento, processamento e análise sejam realizados de forma transparente e em conformidade com padrões legais e morais. De acordo com o Gartner, trata-se de "um sistema de valores e princípios morais relacionados à coleta, ao uso e ao compartilhamento responsáveis de dados", com foco em todas as fases do ciclo de vida dos dados, desde sua geração até sua disseminação. No contexto de projetos baseados em Big Data, como o desenvolvido para a CPTM, a aplicação de padrões éticos é essencial para evitar prejuízos aos usuários e à sociedade, como a perpetuação de desigualdades e violações de privacidade. Assim, a ética deve ser incorporada desde a coleta dos dados até sua análise e apresentação.

4.5. Viés e Discriminação

O tratamento de dados em projetos de grande escala apresenta o risco de introduzir ou perpetuar vieses e discriminações. Esses problemas podem surgir de várias fontes, como dados históricos que carregam desigualdades sociais, processos de coleta enviesados, algoritmos com parâmetros inadequados ou até mesmo a falta de diversidade nas equipes responsáveis pelo desenvolvimento das soluções. Tais fatores, quando negligenciados, podem comprometer a justiça nas decisões, reforçar desigualdades e gerar impactos negativos para grupos sociais menos favorecidos.

Um exemplo prático de viés em um projeto como este seria a priorização de linhas com maior fluxo econômico, em detrimento de regiões periféricas que também enfrentam problemas críticos de transporte. Essa situação pode ocorrer quando os dados utilizados refletem apenas uma parte da realidade, deixando de lado as necessidades de áreas menos favorecidas. Além disso, falhas na modelagem de algoritmos podem resultar em análises que beneficiam desproporcionalmente certos grupos, intensificando a exclusão social.

Para mitigar esses problemas, algumas estratégias podem ser implementadas no projeto. Uma delas é a realização de auditorias frequentes nos dados, verificando sua representatividade em relação à diversidade de usuários do sistema ferroviário. É fundamental assegurar que o conjunto de dados reflita diferentes perfis demográficos, socioeconômicos e geográficos.

Outro aspecto crucial é o treinamento contínuo da equipe de desenvolvimento. É importante que todos os profissionais envolvidos tenham conhecimento sobre os riscos de viés e discriminação, além de compreenderem o impacto social das decisões baseadas em dados. Workshops e cursos sobre ética em ciência de dados, justiça algorítmica e design inclusivo podem ajudar a fortalecer essa perspectiva dentro do projeto.

Testes robustos também são indispensáveis para garantir que os algoritmos desenvolvidos funcionem de maneira justa em diferentes cenários. Por exemplo, ao definir prioridades para reparos em falhas ou alocação de recursos em horários de pico, os modelos devem ser avaliados considerando as demandas de diferentes grupos de passageiros, incluindo aqueles em situações de vulnerabilidade. Esse processo ajuda a evitar que decisões automatizadas beneficiem exclusivamente regiões ou populações específicas.

Por fim, a transparência em todo o processo de análise de dados é essencial para construir confiança. Relatórios claros e acessíveis devem ser disponibilizados, detalhando como os dados foram coletados, tratados e utilizados. Essa prática permite que stakeholders e a sociedade compreendam e questionem as escolhas feitas, promovendo responsabilidade e incentivando melhorias contínuas.

Em suma, abordar o viés e a discriminação em projetos de dados é uma tarefa que exige atenção técnica e compromisso ético. Ao adotar práticas inclusivas e ferramentas de detecção de viés, esse projeto pode garantir que as análises realizadas contribuam para decisões mais justas e impactem positivamente todos os usuários do sistema ferroviário. O combate a essas questões reforça não apenas a qualidade do trabalho, mas também seu alinhamento com os valores sociais de equidade e justiça.

4.6. Responsabilidade social

A responsabilidade social em projetos que utilizam Big Data ultrapassa as obrigações legais, incorporando um compromisso ético com a geração de impactos positivos para a sociedade. No contexto da CPTM, esse compromisso se traduz em tomar decisões informadas a partir dos dados coletados, promovendo melhorias diretas na qualidade de vida dos usuários e otimizando recursos em áreas que mais necessitam de atenção.

Um exemplo claro de responsabilidade social no projeto é a aplicação de análises para identificar horários e estações mais críticos. Essas informações permitem intervenções direcionadas, como aumentar a frequência de trens em momentos de pico ou melhorar a infraestrutura em estações com maior fluxo de passageiros. Além disso, o uso de tecnologias sustentáveis no armazenamento e processamento de dados reduz o impacto ambiental, alinhando o projeto aos princípios de desenvolvimento sustentável.

Outro aspecto relevante é o compromisso com a equidade. As decisões baseadas em dados devem priorizar a redução de desigualdades, direcionando recursos para áreas mais vulneráveis e garantindo que o sistema atenda de forma justa às diversas necessidades da população. Por exemplo, ao identificar regiões menos favorecidas com altos índices de demanda por transporte, a CPTM pode alocar mais trens ou implementar melhorias específicas, fortalecendo sua relação com a comunidade.

Além disso, o projeto contribui para a sociedade ao equilibrar eficiência operacional e impacto social. Ao implementar práticas que garantem decisões éticas e inclusivas, o projeto promove um transporte público mais justo, eficiente e sustentável. A coleta, análise e uso responsável dos dados fortalecem a credibilidade da CPTM e criam uma base sólida para o desenvolvimento contínuo.

Ao incorporar a responsabilidade social como um princípio norteador, o projeto vai além da análise de dados, posicionando-se como uma iniciativa que promove mudanças reais e duradouras. Essa abordagem assegura que os benefícios do projeto sejam compartilhados equitativamente, contribuindo para uma sociedade mais inclusiva e sustentável, enquanto reforça o nosso papel e o papel da CPTM como um exemplo de inovação ética e responsabilidade social.

Conclusão

A implantação do Big Data na CPTM representa um avanço importante, não só para melhorar a operação dos trens, mas também para fortalecer seu compromisso com sustentabilidade, inclusão e transparência. Ao trabalhar com indicadores éticos como privacidade, justiça, responsabilidade social e alinhamento à LGPD, o projeto vai além da tecnologia, focando no impacto positivo para a população e no uso consciente de recursos. Com metas claras e monitoramento constante, a CPTM reforça seu papel como referência em mobilidade urbana responsável, garantindo que os benefícios cheguem de forma justa a todos os usuários.

5. Streamlit e Infográfico

Essa seção é dedicada à documentação do que temos no nosso DataApp: o código **Streamlit** e o **Infográfico**, na última página do front-end.

5.1. Documentação do Streamlit

O Streamlit é uma biblioteca de código aberto em Python projetada para simplificar o desenvolvimento de aplicações web interativas e personalizáveis, voltadas para a visualização e exploração de dados. Criado em 2019, o Streamlit se destaca por sua abordagem intuitiva e minimalista, permitindo que cientistas de dados, analistas e desenvolvedores criem rapidamente interfaces gráficas sem necessidade de conhecimentos avançados em desenvolvimento web. ([Streamlit Documentation, 2024](#))

A principal vantagem do Streamlit é sua integração nativa com bibliotecas de ciência de dados populares, como pandas, NumPy, Matplotlib e Plotly, tornando-o uma escolha ideal para prototipagem rápida e compartilhamento de insights. Com comandos simples e um foco na produtividade, ele transforma scripts de Python em aplicativos web interativos executados localmente ou na nuvem em uma velocidade fora do normal.

O Streamlit foi utilizado como ferramenta para desenvolver um dashboard interativo que apresenta os dados da CPTM de forma visual e acessível. Com ele, os dados são processados e exibidos dinamicamente, permitindo que os usuários naveguem por diferentes páginas e explorem insights relacionados à operação ferroviária e ao comportamento dos passageiros.

5.1.1. Autenticação

Antes de o usuário chegar ao dashboard principal, é necessário passar por uma autenticação via login e senha. Esse mecanismo garante que apenas pessoas autorizadas accessem as informações operacionais e estratégicas, mantendo a confidencialidade dos dados e reforçando a credibilidade das análises. Ao fornecer suas credenciais, o usuário é identificado, o que facilita a rastreabilidade de suas ações, a auditoria interna e o monitoramento da utilização do sistema.

A experiência de login é simples e direta: ao acessar a plataforma, o usuário é imediatamente direcionado para a tela de autenticação, onde insere nome de usuário e senha. Caso as credenciais sejam válidas, a sessão é marcada como autenticada, liberando o acesso ao dashboard completo. Em caso de falha, uma mensagem orienta o usuário a verificar suas informações, assegurando assim que o acesso seja restrito apenas a indivíduos devidamente credenciados.

Na imagem abaixo, é possível visualizar a tela de login, que solicita ao usuário suas credenciais, garantindo que somente indivíduos autorizados accessem a solução.

Figura 14 - Tela de Autenticação



Fonte: Leandro Carvalho (2024)

Após o login bem-sucedido, o usuário é redirecionado ao dashboard principal, onde pode explorar métricas, relatórios e demais dados operacionais e estratégicos de forma segura, como pode ser visto na imagem abaixo:

Figura 15 - Template Data Product Canvas

A screenshot of a dark-themed dashboard titled "Visão Estratégica - CPTM". The sidebar on the left lists various operational metrics: Home, Fluxo Entre Estações, Heatmap, Intervalo Médio Operação, Tempo Porta Aberta, Movimento Classificado, Tipos de Bilhete, Sensores por Data, and Infográfico. A green notification bar at the top right says "Login bem-sucedido!". The main content area displays a title "Visão Estratégica - CPTM" with a small train icon, and a message "Bem-vindo ao painel estratégico. Explore os dados operacionais e filtre informações relevantes para a tomada de decisão." The top right corner shows deployment status: "RUNNING...", "Stop", "Deploy", and a zoom control "- 125% + Redefinir".

Fonte: Leandro Carvalho (2024)

Essas imagens ilustram o fluxo básico do usuário, desde o acesso inicial, passando pela autenticação, até a navegação no painel principal. Dessa forma, a camada de segurança não apenas protege a informação, mas também assegura uma experiência de uso estruturada e confiável.

5.1.2. Dashboard

A aplicação desenvolvida com Streamlit é estruturada em páginas, cada uma dedicada a um conjunto específico de análises e visualizações. Cada página obtém seus dados por meio de chamadas a endpoints GET, que interagem com a API Flask. Essa API atua como intermediária entre o dashboard e o banco de dados, garantindo segurança, atualizações constantes e integridade das informações apresentadas. Assim, o usuário tem sempre à disposição dados atualizados, pois o Streamlit executa as funções de coleta toda vez que uma página é carregada ou atualizada.

A estrutura das páginas é a seguinte:

Figura 16 - DataApp - Estrutura das Páginas



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

1. **Home:** Página inicial com uma introdução e navegação para as demais análises.
2. **Fluxo Entre Estações:** Exibe o fluxo de entrada e saída de passageiros em diferentes estações.
3. **Heatmap:** Apresenta um heatmap de movimentações de passageiros por linha e horário.
4. **Intervalo Médio Operação:** Mostra o intervalo médio de operação ao longo do dia.
5. **Tempo Porta Aberta:** Analisa o tempo médio em que as portas dos trens permanecem abertas.
6. **Movimento Classificado:** Classifica os movimentos por tipo de bilhete.
7. **Tipos de Bilhete:** Apresenta os tipos de bilhetes mais utilizados e sua distribuição ao longo do tempo.
8. **Infográfico:** Essa é uma página dedicada somente ao Infográfico desenvolvido em uma aula de UX, que foi explicado na próxima seção da documentação.

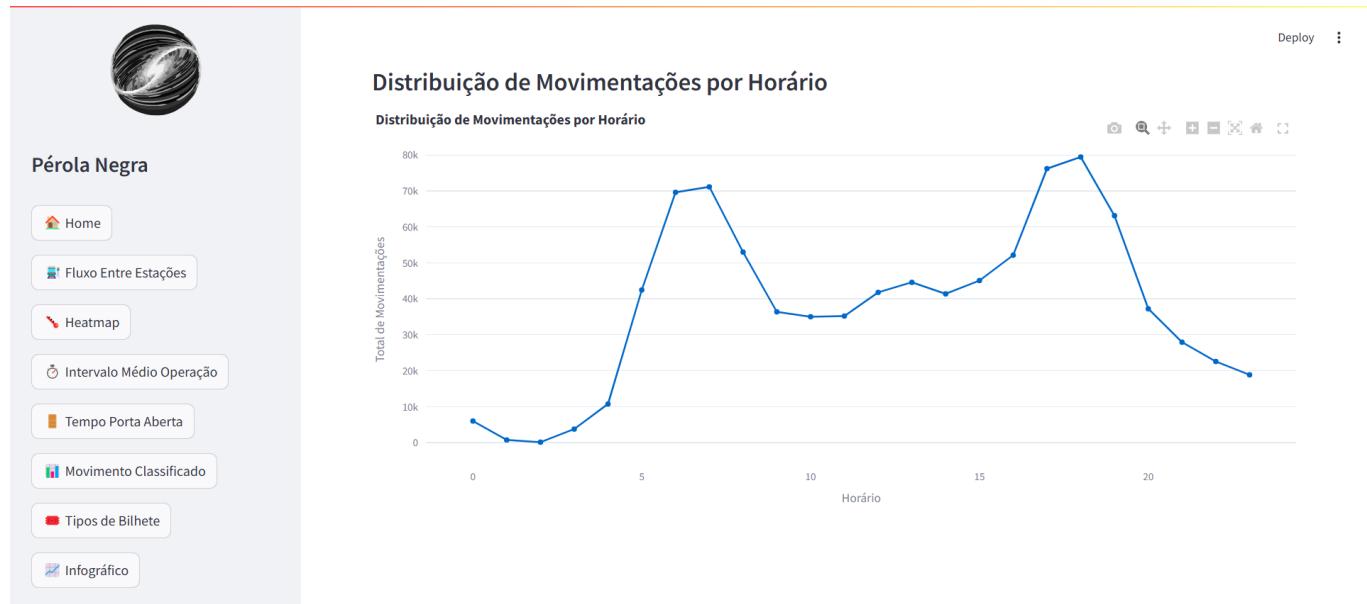
Toda vez que uma página é carregada, o Streamlit executa as funções associadas para realizar chamadas aos endpoints de GET. Isso significa que o dashboard está sempre atualizado com as informações mais recentes disponíveis no banco de dados.

As funções abaixo foram implementadas para coletar os dados necessários e alimentar cada página do dashboard:

- **get_fluxo_entre_estacoes:** Obtém o fluxo de passageiros entre estações.
- **get_heatmap_pessoas_por_linha:** Coleta dados para gerar um heatmap de movimentações por linha e horário.
- **get_media_intervalo_operacao_por_dia:** Recupera dados sobre o intervalo médio de operação durante o dia.
- **get_media_tempo_porta_aberta:** Calcula o tempo médio de porta aberta dos trens.
- **get_movimento_classificado_por_bilhete:** Classifica os movimentos dos passageiros por tipo de bilhete.
- **get_tipos_bilhete_abundantes:** Identifica os bilhetes mais utilizados.
- **get_tipos_bilhete_por_dia:** Analisa o uso de bilhetes por dia.
- **get_tipos_bilhete_por_semana:** Extrai dados semanais sobre os bilhetes utilizados.

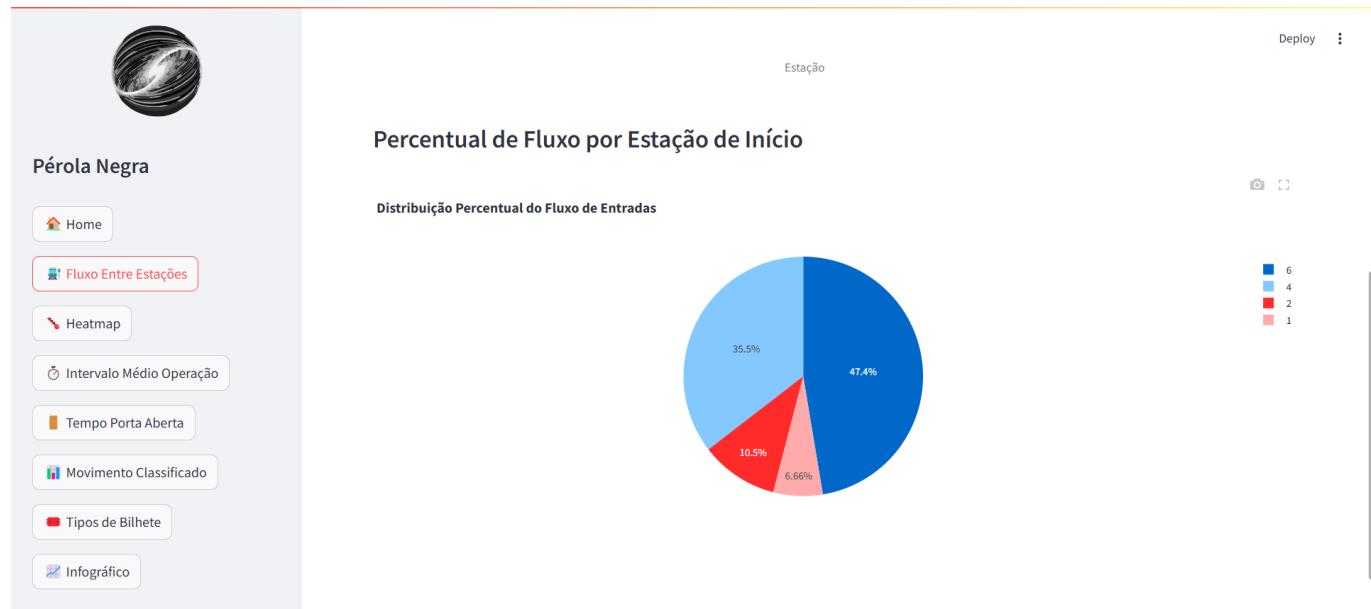
ANas imagens abaixo é possível visualizar como está a primeira versão do DataApp.

Figura 17 - DataApp - Distribuição de Movimentações por Horário



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 18 - DataApp - Fluxo por Estação



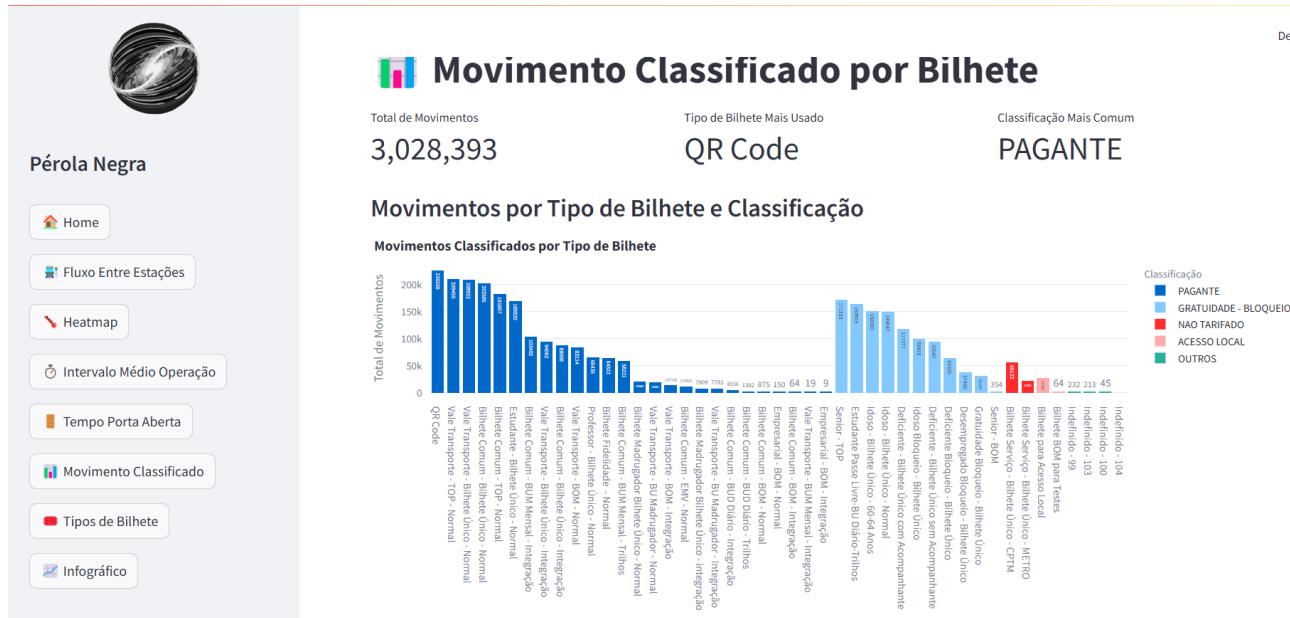
Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 19 - DataApp - Tempo Médio de Porta Aberta



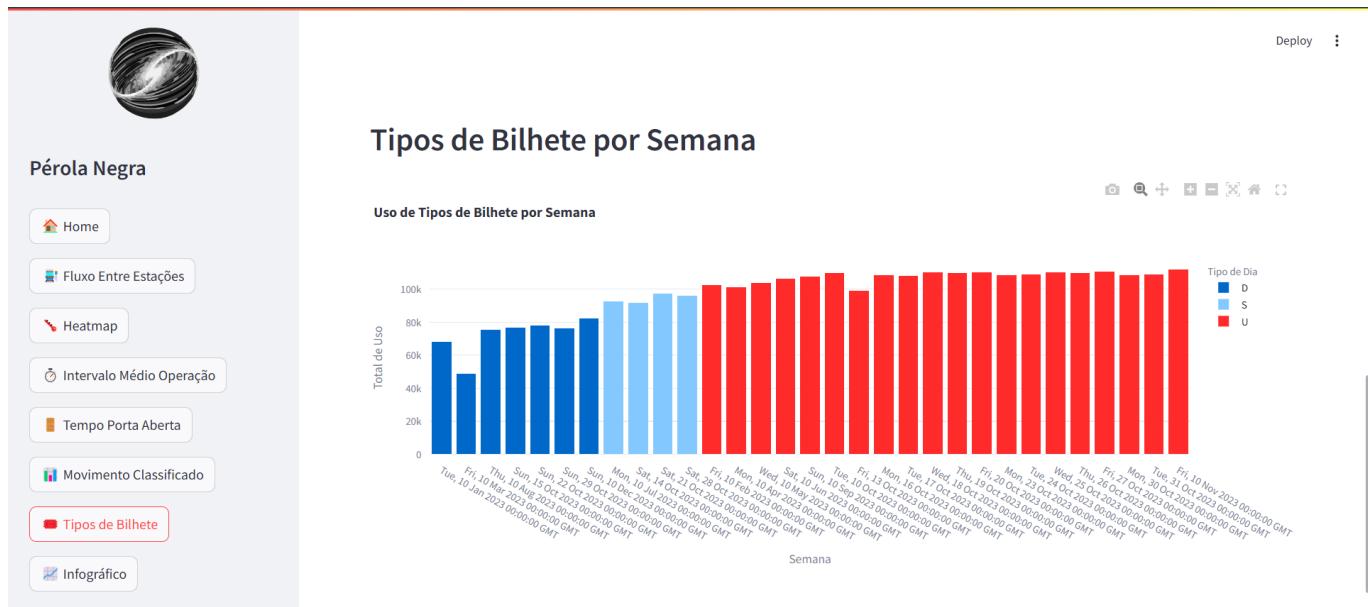
Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 20 - DataApp - Movimento Classificado por Bilhete



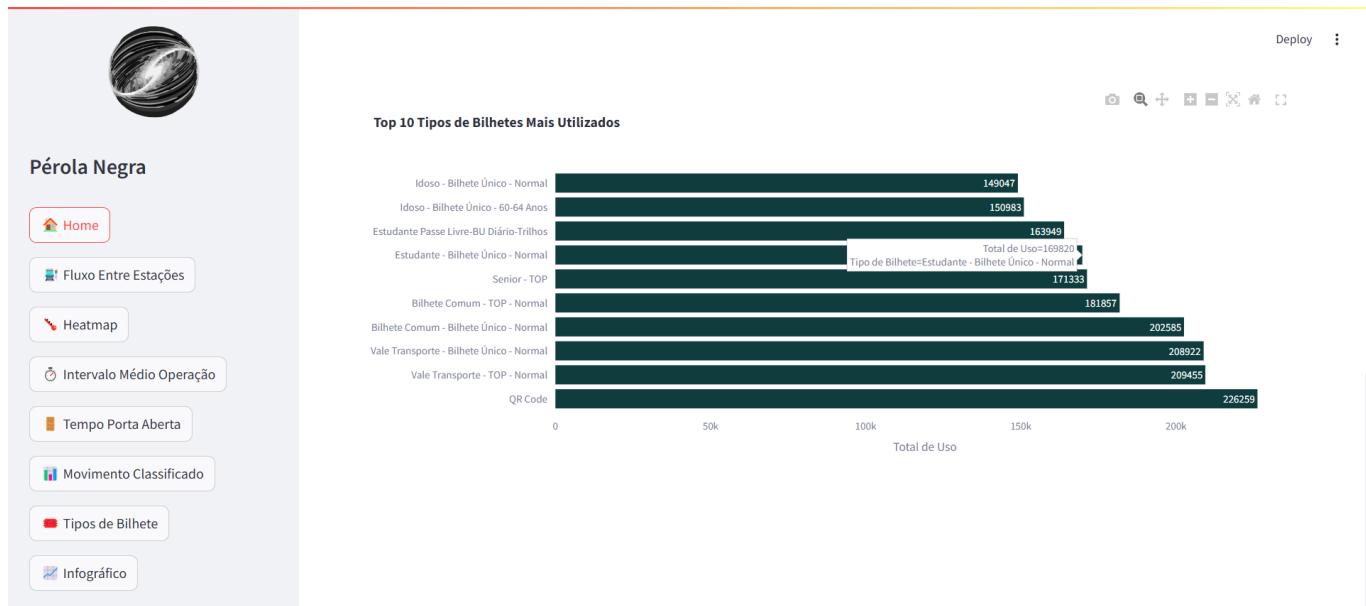
Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 21 - DataApp - Tipos de Bilhete por Semana



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 22 - DataApp - Top 10 Tipos de Bilhetes Mais Utilizados



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

O uso do Streamlit no projeto proporciona uma experiência interativa e dinâmica para a visualização de dados da CPTM. Com a atualização automática dos endpoints de GET a cada recarregamento de página, o dashboard garante que os dados exibidos estejam sempre atualizados, oferecendo uma base confiável para análises e tomadas de decisão.

5.2. Documentação dos Filtros

Além de todos os gráficos, também foram criados filtros para alguns deles, os quais estão descritos logo abaixo com imagens. Esses filtros servem para ajudar o usuário a encontrar mais rapidamente as informações específicas que procura.

Figura 23 - DataApp - Heatmap - Seleção de Linhas

The interface features a title "Heatmap de Movimentação de Pessoas" with a red key icon. Below it is a dropdown menu labeled "Selecionar a linha (ou 'Todas' para todas as linhas):" containing the option "Todas". A scrollable list of line numbers follows:

- Todas
- 13
- 97
- 98
- 99
- 100

At the bottom of the list are page navigation controls: "2" and "3".

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico "Heatmap de Movimentação", o filtro de seleção de linha permite que o usuário escolha uma linha específica ou opte pela opção "Todas", incluindo assim todo o conjunto. Esse filtro traz flexibilidade ao analisar a movimentação dos passageiros, tornando possível focar em uma linha de interesse para entender padrões de uso, identificar horários de pico e necessidades operacionais relacionadas a determinadas rotas. Ao mesmo tempo, a opção de "Todas" facilita uma visão panorâmica de todas as linhas, auxiliando na comparação geral do desempenho e no planejamento estratégico.

Figura 24 - DataApp - Filtro - Heatmap Intervalo de Horários

Heatmap de Movimentação de

Selecione a linha (ou 'Todas' para todas as linhas):

13

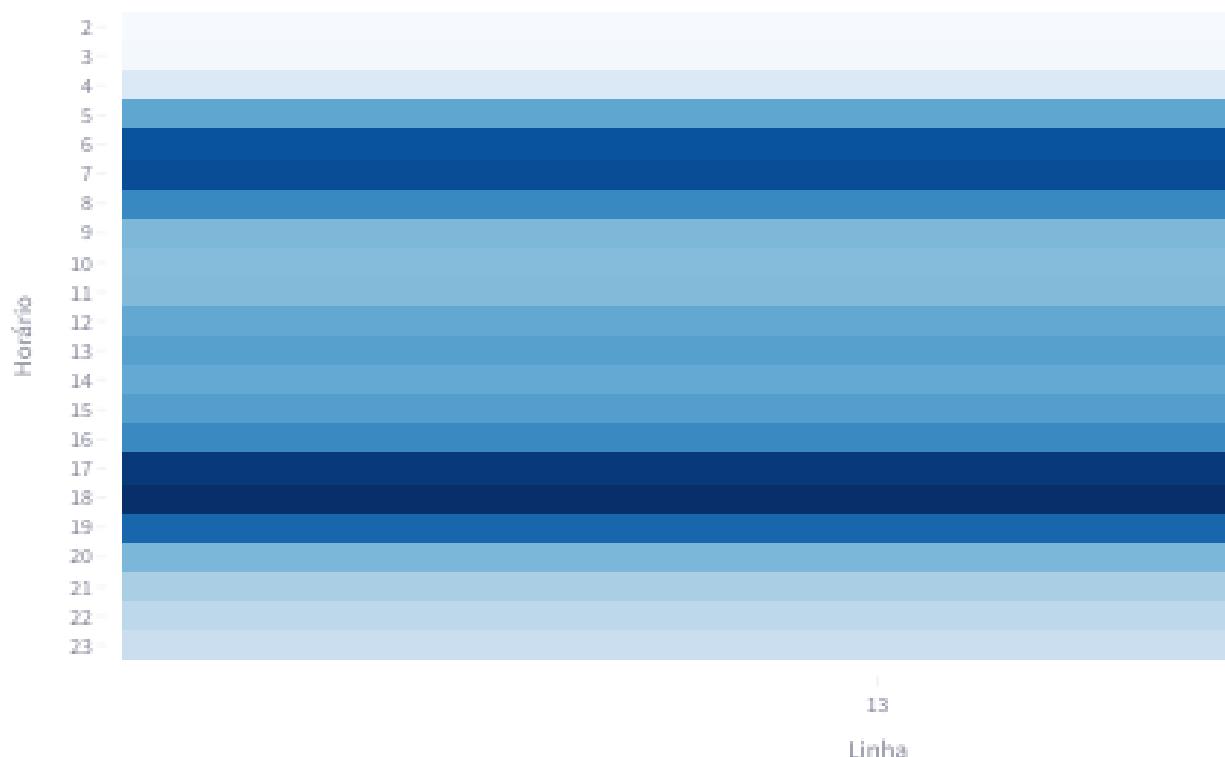
Selecione o intervalo de horários:

0 2



Mapa de Calor - Movimentação por Linha e Horário

Heatmap - Movi



Total Movimentações

890,773 pessoas

Linha Mais Movimentada

Linha 13

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Após selecionar a linha, como mostrado na figura 24, o filtro torna possível a seleção de intervalo de horário, por meio de uma barra vermelha deslizante. Esse recurso permite que o usuário limite a análise a um período específico do dia, seja o horário de pico da manhã, o final da tarde ou um intervalo pré-determinado. Ao ajustar o intervalo, o usuário pode detectar padrões temporais de movimentação, identificar gargalos nos horários de maior demanda e ajustar recursos, como número de trens ou funcionários, para melhorar a eficiência operacional.

Figura 25 - DataApp - Filtro - Eventos Críticos por Data

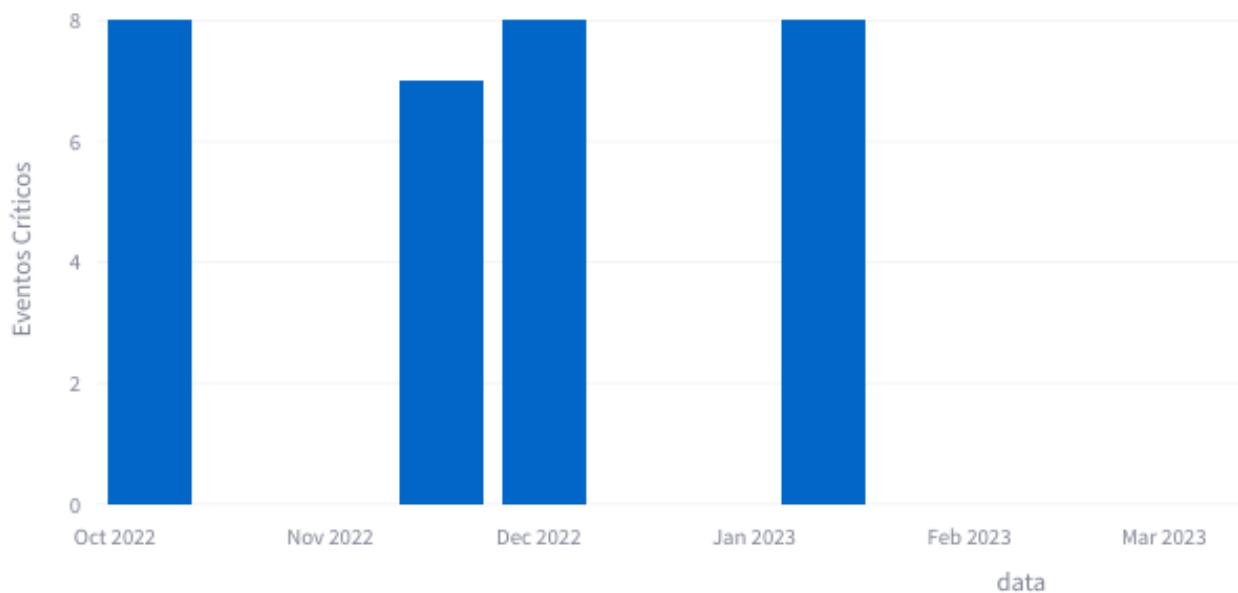
Eventos Críticos por Data

Intervalo de datas para o Gráfico 4

2020-09-17



Eventos Críticos por Data



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico "Eventos Críticos por Data", o filtro de intervalo de datas possibilita que o usuário selecione um período específico para análise. Essa funcionalidade é essencial para investigações temporais, permitindo observar se houve aumento de eventos críticos em certos meses, ou a queda após iniciativas de manutenção. Ao restringir o período, é mais fácil correlacionar os incidentes com fatores externos (como clima, feriados ou eventos na cidade) e tomar decisões informadas sobre alocações de recursos e ações preventivas.

Figura 26 - DataApp - Filtro - Tendência de Eventos ao Longo do Tempo

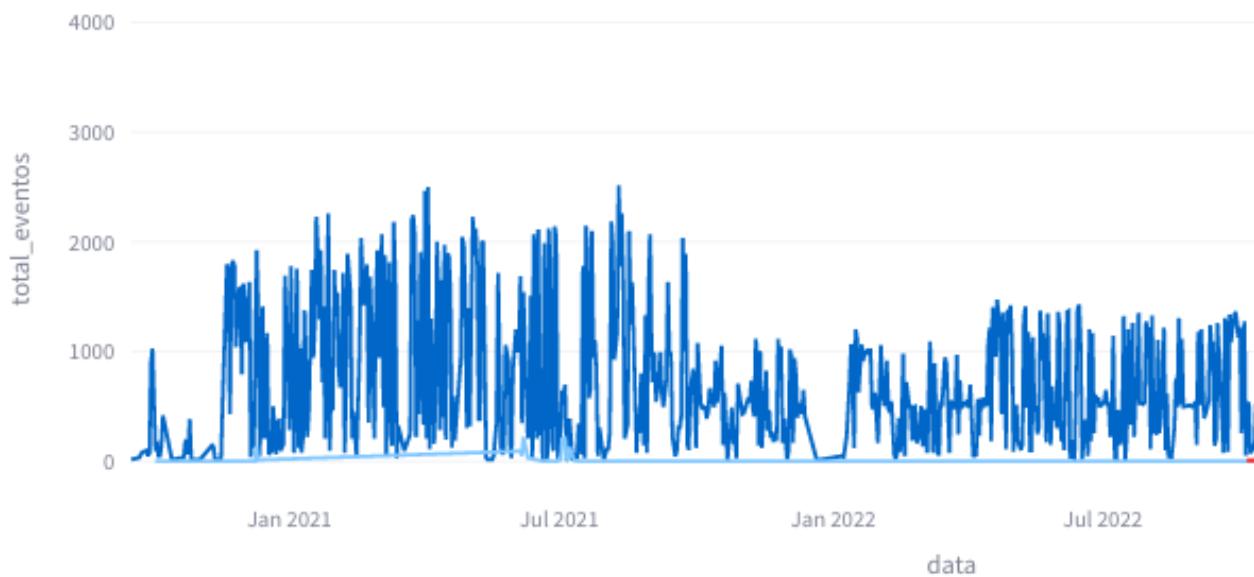
Tendência de Eventos ao Longo do Tempo

Intervalo de datas para o Gráfico 3

2020-09-17

2020-09-17

Tendência de Eventos por Status



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico “Tendência de Eventos ao Longo do Tempo”, o filtro de datas atua como um zoom temporal, permitindo avaliar tendências e flutuações em períodos específicos. Ajustando esse intervalo, o usuário pode analisar a evolução dos eventos ao longo dos anos, verificar se as medidas corretivas implementadas surtiram efeito e identificar padrões de sazonalidade. Essa visão refinada apoia o planejamento de longo prazo e a melhoria contínua do serviço.

Figura 27 - DataApp - Filtro - Proporção de Status dos Sensores

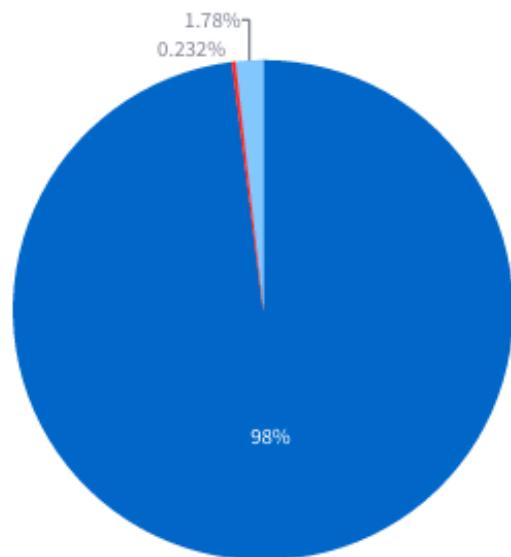
Proporção de Status dos Sensores

Intervalo de datas para o Gráfico 2

2020-09-17

2020-09-17

Distribuição Percentual dos Status



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico “Proporção de Status dos Sensores”, o filtro de datas delimita a janela de observação para analisar a condição dos sensores em determinado intervalo. Desse modo, é possível verificar se houve variações significativas nas proporções de status em períodos específicos. Essa informação é útil para monitorar a eficácia da manutenção preventiva, avaliar a estabilidade do sistema de sensores e antecipar intervenções técnicas antes que um volume maior de falhas ocorra.

Figura 28 - DataApp - Filtro - Total de Eventos por Data por Linha

Total de Eventos por Data (Linha Selecionada)

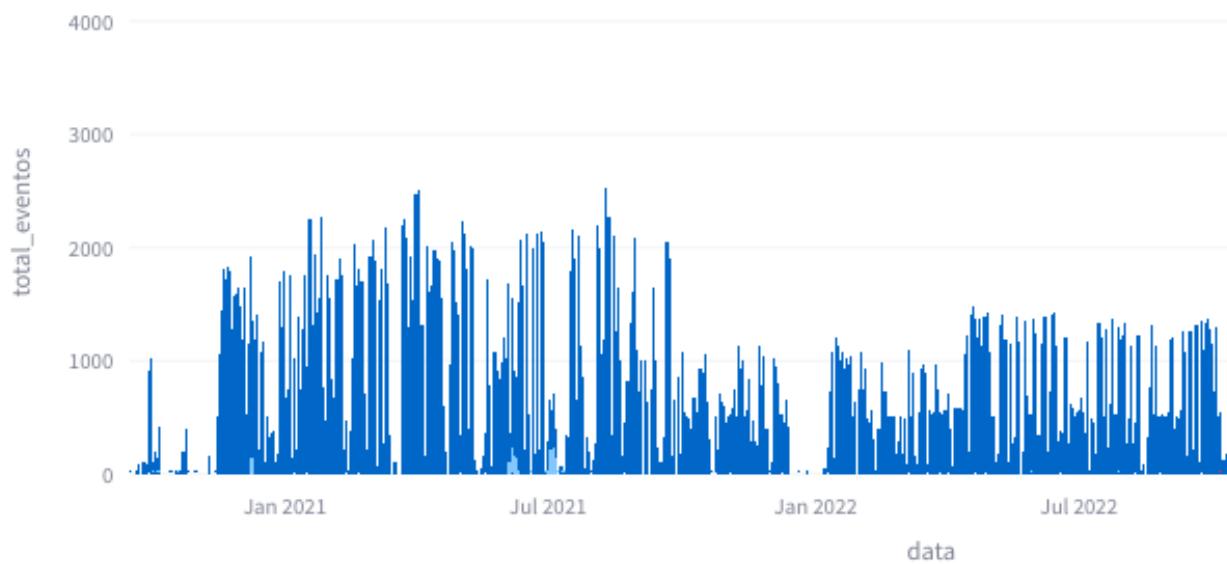
Intervalo de datas para o Gráfico 1

2020-09-17



2020-09-17

Eventos por Status - Linha Todos



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No gráfico “Total de Eventos por Data (Linha Selecionada)”, o filtro de datas combinado à seleção de linha permite entender a dinâmica de eventos ao longo do tempo em um contexto mais restrito. Ajustando o intervalo de datas, o usuário pode analisar períodos críticos, como semanas de manutenção intensiva ou temporadas com demanda atípica, e avaliar o impacto em linhas específicas. Essa abordagem facilita a correção de falhas, a otimização de rotas e a melhoria na alocação de recursos.

Figura 29 - DataApp - Filtro - Sensores por Data

Sensores por Data

Selecione uma Linha:

Todos

OK

Warning

3800

69

Total de Eventos por Data (Linha Selecionada)

Intervalo de datas para o Gráfico 1

2020-09-17

2020-09-17

Fonte: Material produzido pelo Grupo Pérola Negra (2024)

No painel “Sensores por Data”, o filtro de seleção de linha permite que o usuário escolha uma linha específica ou mantenha a opção “Todos”. Ao focar em uma linha específica, é possível detectar problemas pontuais de sensores naquele trajeto, entender padrões de falhas e necessidades de manutenção mais frequentes. Já a opção “Todos” oferece uma visão geral, possibilitando a comparação entre diferentes linhas e auxiliando na priorização de investimentos em tecnologia e manutenção.

Figura 30 - DataApp - Filtro - Entradas e Saídas por Estação

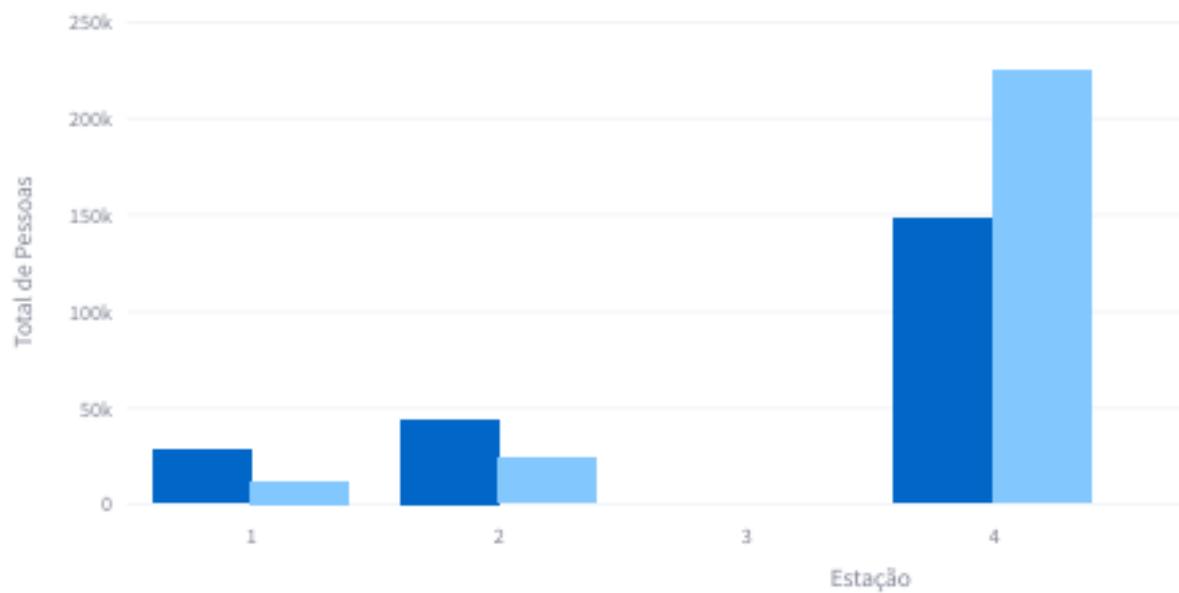
Fluxo Entre Estações

Selecione as estações:

Todos ×

Gráfico de Entradas e Saídas por Estação

Comparativo de Entradas e Saídas em todas as datas



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

O filtro acima, utilizado no gráfico "Fluxo Entre Estações," é responsável por permitir que o usuário escolha quais estações deseja visualizar, além de incluir opções para selecionar todas ou nenhuma estação. Esse filtro é essencial para personalizar a análise, permitindo foco em estações específicas de interesse ou uma visão ampla do fluxo entre todas as estações. Ele facilita a identificação de padrões ou anomalias em determinadas estações, auxiliando na tomada de decisões baseadas em dados operacionais.

Figura 31 - DataApp - Filtro - Tempo Médio de Porta Aberta por Linha

Tempo Médio de Porta Aberta

Data de início:

2020/09/17

Data de fim:

2023/09/23

Resumo do Tempo Médio de Porta Aberta por Linha

Linha 13.0 ⓘ

13.62 min

↑ 10.11 min

Linha 97.0 ⓘ

3.65 min

↑ 20.08 min

Linha 98.0 ⓘ

48.64 min

↑ 24.91 min

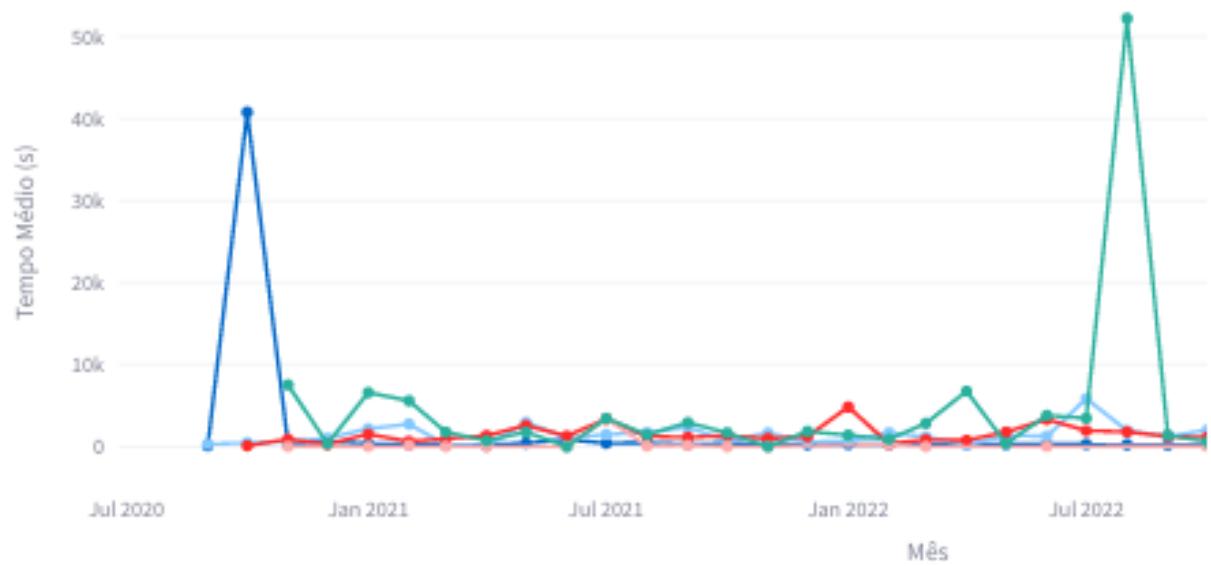
Linha

25

↑ 1

Variação Mensal do Tempo Médio de Porta Aberta por Linha

Variação Mensal do Tempo Médio de Porta Aberta



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

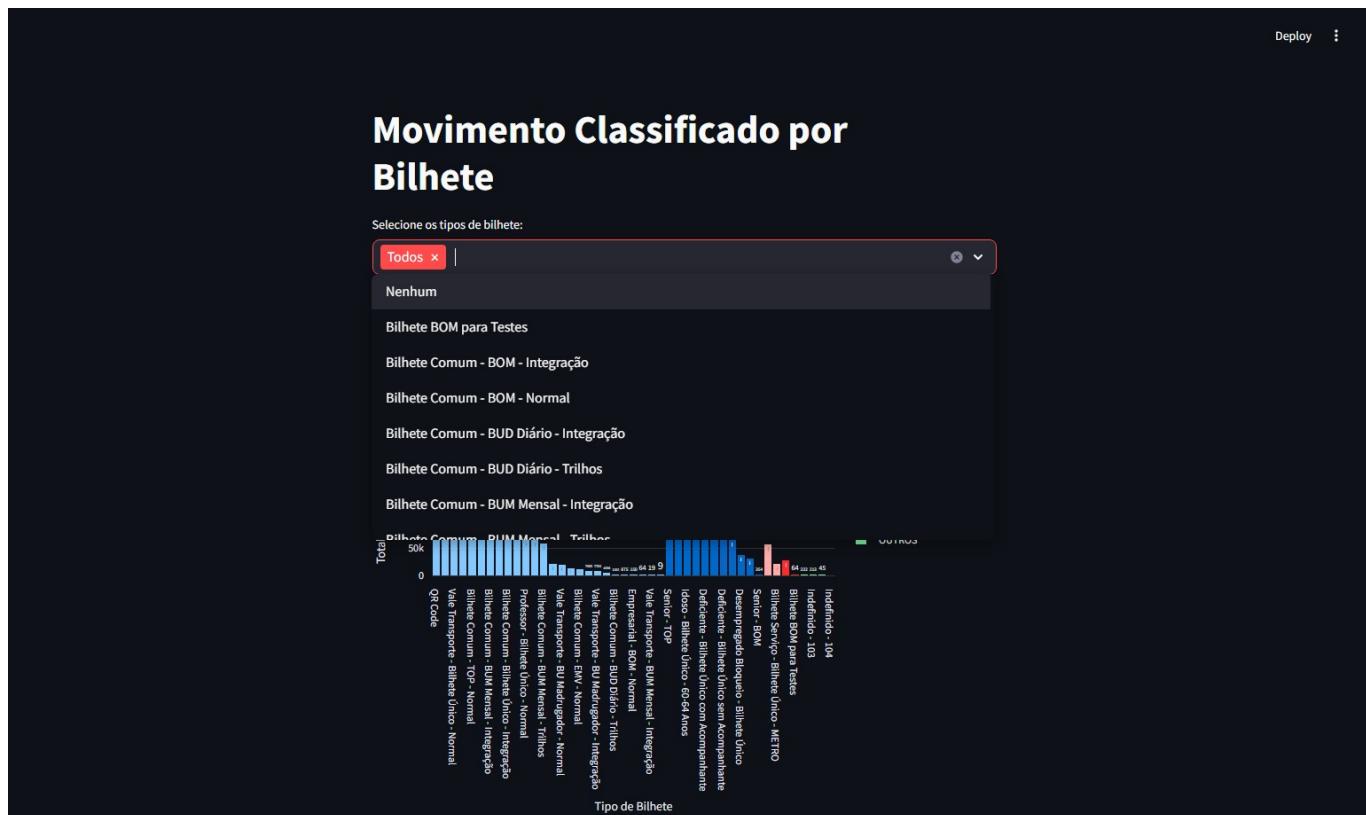
Na visualização do "Tempo Médio de Porta Aberta", o usuário pode ajustar as datas de início e fim, delimitando o período a ser analisado. Dessa forma, é possível observar se o tempo médio de porta aberta aumentou ou diminuiu após certa intervenção, se determinados meses são mais críticos devido ao clima ou horários de pico, e quais linhas apresentam maior variação ao longo de um intervalo. Esse recurso garante uma análise mais precisa e contextualizada, auxiliando no aprimoramento da operação.

Figura 32 - DataApp - Filtro - Movimento Classificado por Bilhete 1



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Figura 33 - DataApp - Filtro - Movimento Classificado por Bilhete 2

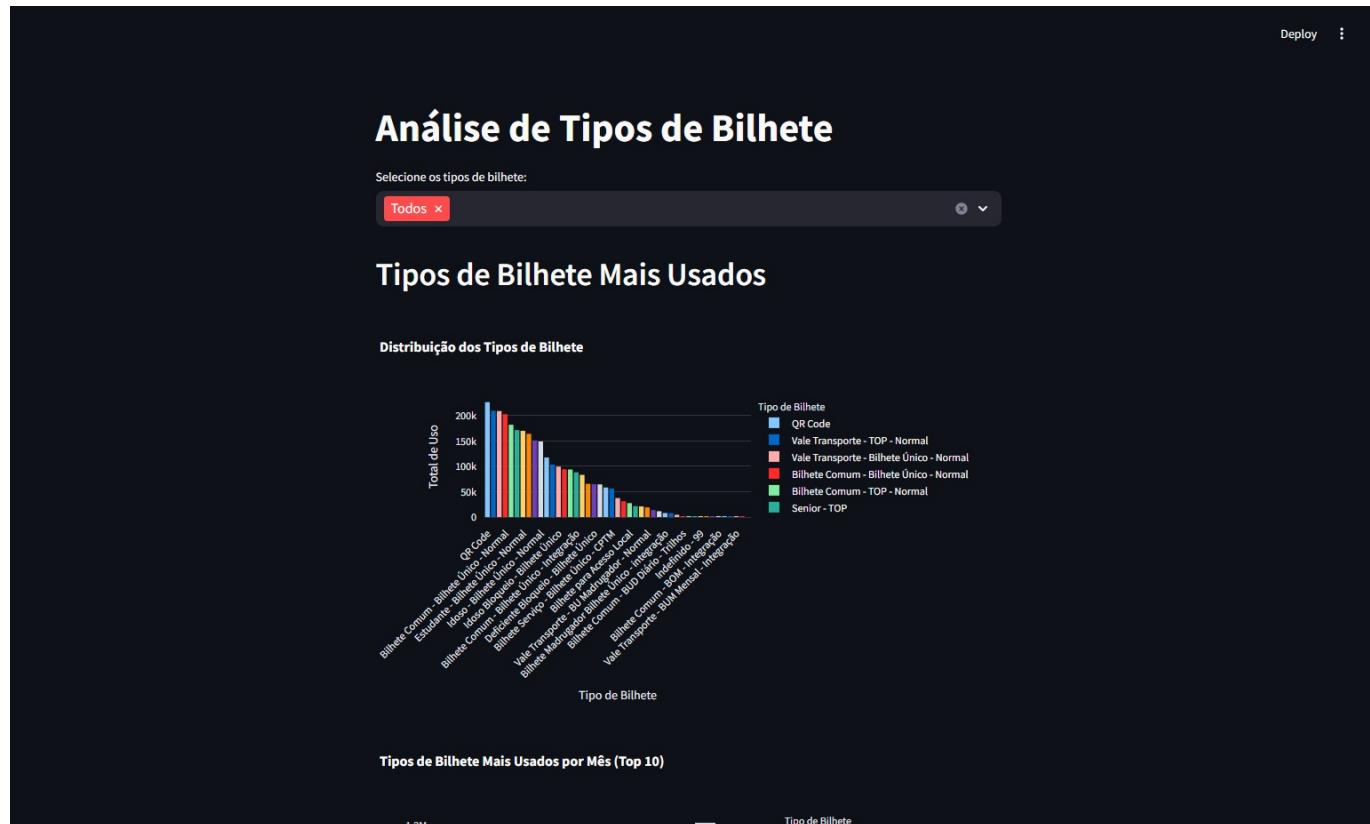


Fonte: Material produzido pelo Grupo Pérola Negra (2024)

O filtro acima, do gráfico "Movimento Classificado por Bilhete", tem uma função parecida com o primeiro filtro. Ele permite que o usuário selecione o bilhete que deseja visualizar, além de oferecer as opções de exibir

todos ou nenhum bilhete. Esse filtro é importante para análises específicas por tipo de bilhete, ajudando a identificar tendências no uso de categorias específicas, como bilhetes de estudante, VT ou QR Code. Além disso, o gráfico associado oferece a funcionalidade de zoom, permitindo que o usuário amplie partes específicas, como dias da semana, para uma análise mais detalhada e segmentada, fornecendo insights valiosos para a otimização do sistema de bilhetagem e o atendimento aos passageiros.

Figura 34 - DataApp - Filtro - Análise de Tipos de Bilhete



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Por fim, o filtro acima (de Análise de Tipos de Bilhete) é extremamente parecido com o anterior, pois também faz uma seleção de bilhetes. Porém, ele se limita a mostrar um gráfico de barras com os tipos de bilhetes mais utilizados, sem considerar dados específicos por dia da semana ou períodos, dado que é mostrado em outro gráfico. Essa simplicidade torna o filtro eficiente para identificar rapidamente os bilhetes mais populares, auxiliando na priorização de estratégias focadas nas categorias mais utilizadas, como melhorias no atendimento ou campanhas promocionais direcionadas.

5.3. Documentação do Infográfico

Como foi dito na seção anterior relacionada ao código do Streamlit, a última página do DataApp foi dedicada para mostrar um infográfico. Ele foi criado para apresentar uma retrospectiva histórica e destacar dados relevantes sobre o uso dos trens e bilhetes da CPTM. Ele combina um storytelling visual sobre a evolução dos trilhos e os tipos de bilhetes mais utilizados atualmente, permitindo que os usuários explorem essas informações de forma interativa. A ideia principal é conectar o impacto histórico dos trens ao comportamento contemporâneo de seus usuários.

Abaixo pode-se conferir uma imagem do mesmo.

Figura 35 - Infográfico

Dos Trilhos à Rotina

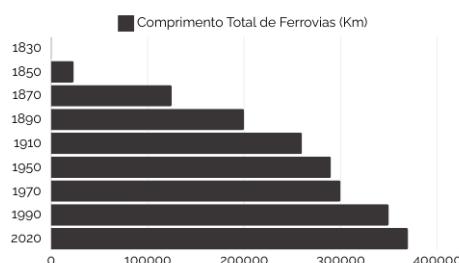
A Evolução dos Trens como Meio Essencial para Trabalhadores

A Revolução Industrial e os Trilhos

"A História na Europa"

Século XIX

No século XIX, as ferrovias impulsionaram a Revolução Industrial na Europa, facilitando o transporte de matérias-primas, mercadorias e trabalhadores para as fábricas nos centros urbanos.



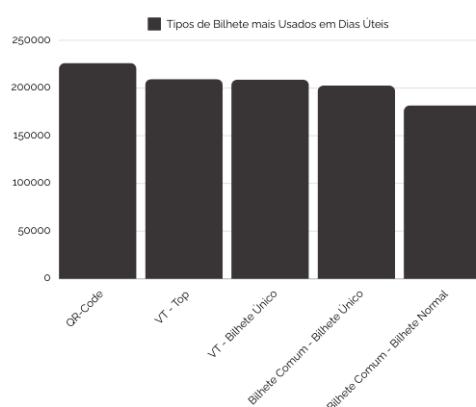
Com esse gráfico, é possível observar que houve um salto consideravelmente grande na construção de trilhos ferroviários de 1850 para 1870 e de 1870 para 1890 também, mas depois desses períodos, a construção de trilhos cai cada vez mais em frequência.

Dias Atuais...

"O Papel dos Trens no Brasil Moderno"

A CPTM e os Trabalhadores

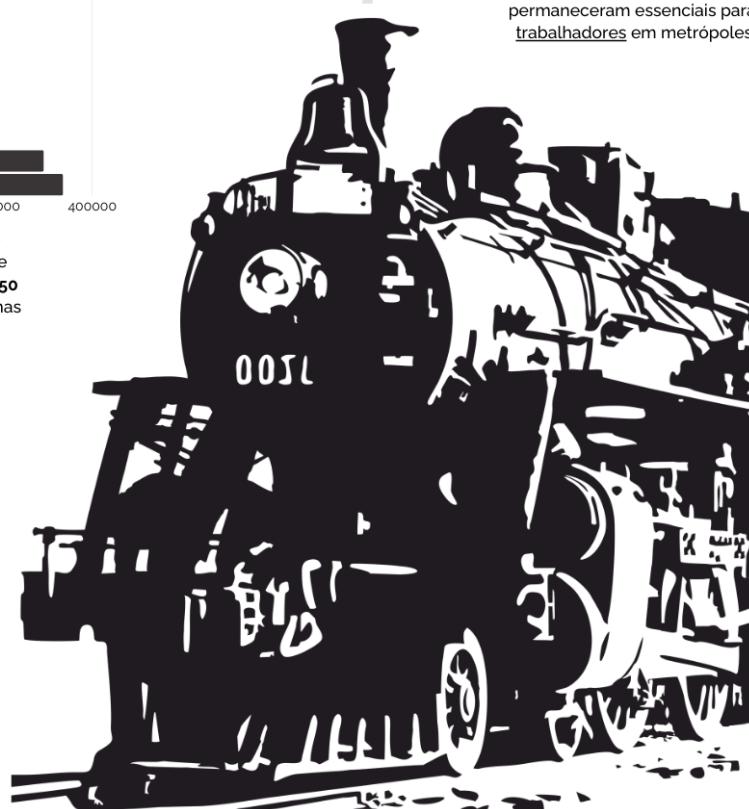
Dados recentes mostram como os trens são utilizados pela população. Neles, é mostrado que a maioria dos bilhetes ainda são comprados na hora, mas que o Vale Transporte (tanto bilhete único, quanto TOP) que é fornecido como benefício para trabalhadores CLT, toma o segundo e terceiro lugar no ranking de tipos de bilhetes comprados. Ou seja, o sistema ferroviário sempre foi e continua sendo um dos meios de transporte mais importantes para trabalhadores.



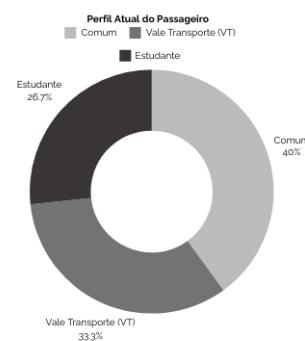
Trilhos no Brasil - 1854

"Da Economia ao Cotidiano"

A Estrada de Ferro Mauá marcou o início das ferrovias no Brasil, conectando interior e litoral. Na Era das Ferrovias (1870-1920), impulsionou o transporte agrícola. Mesmo após a desativação de linhas com o foco rodoviário (1950), os trens urbanos permaneceram essenciais para trabalhadores em metrópoles.



O gráfico de pizza à direita revela que, entre os tipos de bilhete da CPTM, o mais utilizado é Comum, representando 40%. Em segundo lugar está o bilhete Vale Transporte, com 33.3%, enquanto o menos utilizado é Estudante, com 26.7%. Novamente é possível perceber como os trabalhadores utilizam bastante esse meio de transporte.



Fonte: Material produzido pelo Grupo Pérola Negra (2024)

Ele foi projetado com uma estrutura de três partes, sendo elas a "Revolução Industrial e os Trilhos", "Trilhos no Brasil - 1854" e "Dias Atuais". Nessa primeira seção, foi explicada a timeline das rodoviárias na Europa, desde sua criação em 1830, até 2020. Isso é mostrado tanto por um curto texto explicativo, quanto por um gráfico de barras horizontais, mostrando a evolução dos trens durante os últimos séculos.

A segunda seção, por mais que ainda seja sobre o passado, muda de área, saindo da Europa para o Brasil. Nela não há nenhum gráfico, somente um texto explicando que "Mesmo após a desativação de linhas com o foco rodoviário (1950), os trens urbanos permaneceram essenciais para trabalhadores em metrópoles." Dessa forma, conseguimos introduzir o terceiro e último tópico do Infográfico.

Depois de contar a história dos trens na Europa e no Brasil, podemos falar sobre a rodovia nos dias de hoje. Nessa seção há 2 gráficos: um gráfico de rosca e um de barras verticais, ambos contendo informações parecidas, retiradas da base de dados da CPTM. O primeiro gráfico mostra os tipos de bilhetes mais utilizados em dias úteis, e, assim, podemos analisar que há diversos trabalhadores nos trens da CPTM, já que os segundo e terceiro tipos de bilhetes mais usados são tipos de Vale Transporte. Partindo para o último gráfico, ele mostra quais são os usuários que mais possuem bilhetes de passagem dos trens da CPTM. Ao observar que o bilhete Vale Transporte, normalmente utilizado por trabalhadores como benefício da empresa em que atuam, é o 2º mais utilizado, podemos afirmar novamente que há muitos trabalhadores nos trens da CPTM.

5.4. Geração de Relatórios via Botão no Streamlit

Para aprimorar ainda mais o DataApp, foi implementado um sistema de geração automática de relatórios através do dashboard desenvolvido com **Streamlit**. Essa funcionalidade permite que os usuários gerem relatórios das análises visualizadas em cada página da aplicação.

Funcionalidade do Botão de Geração de Relatório

A funcionalidade de geração de relatórios é acionada por um botão presente em cada página do dashboard. Ao clicar no botão, o sistema compila os dados e visualizações atuais da página, formata-os adequadamente e disponibiliza um arquivo em formato txt para download.

Fluxo de Operação

1. **Clique no Botão:** O usuário pressiona o botão "Gerar Relatório" e em seguida no botão "Baixar Relatório".
2. **Coleta de Dados:** O sistema identifica a página atual e coleta os dados exibidos nela.
3. **Formatação dos Dados:** Utilizando funções específicas, os dados são formatados para garantir legibilidade e consistência.
4. **Geração do Relatório:** O relatório é gerado em formato .txt
5. **Disponibilização para Download:** O relatório gerado é disponibilizado para download pelo usuário.

Implementação Técnica

A seguir, apresentamos o código implementado para essa funcionalidade, juntamente com explicações detalhadas de cada parte.

Funções de Formatação e Mapeamento de Páginas

```

def format_number_report(value):
    if value >= 1_000_000:
        return f"{value / 1_000_000:.1f}M"
    elif value >= 1_000:
        return f"{value / 1_000:.1f}k"
    return str(value)

page_to_filename = {
    "🏠 Home": "relatorio_home.txt",
    "👤 Fluxo Entre Estações": "relatorio_fluxo_entre_estacoes.txt",
    "🌡 Heatmap": "relatorio_heatmap.txt",
    "🕒 Intervalo Médio Operação": "relatorio_intervalo_medio_operacao.txt",
    "🕒 Tempo Porta Aberta": "relatorio_tempo_porta_aberta.txt",
    "📊 Movimento Classificado": "relatorio_movimento_classificado.txt",
    "🎫 Tipos de Bilhete": "relatorio_tipos_de_bilhete.txt",
    "🕒 Sensores por Data": "relatorio_sensores_por_data.txt",
    "📈 Infográfico": "relatorio_infografico.txt",
}

def get_report_filename(page):
    return page_to_filename.get(page, "relatorio_analise_dados.txt")

```

- **format_number_report**: Esta função formata números grandes para torná-los mais legíveis, convertendo valores em milhares (**k**) ou milhões (**M**).
- **page_to_filename**: Um dicionário que mapeia cada página do dashboard para criar um nome de arquivo específico para o relatório.
- **get_report_filename**: Função que retorna o nome do arquivo de relatório com base na página atual. Se a página não estiver no dicionário, retorna um nome padrão.

Função de Geração de Relatório

```

def generate_report(page):
    report = "Relatório de Análise de Dados\n"
    report += "="*30 + "\n\n"

    if page == "🏠 Home":
        report += "Página: 🏠 Home\n"
        report += "-"*20 + "\n"

        fluxo_data = get_fluxo_entre_estacoes()
        intervalo_data = get_media_intervalo_operacao_por_dia()
        porta_data = get_media_tempo_porta_aberta()

        if fluxo_data:
            df_fluxo = pd.DataFrame(fluxo_data)
            df_fluxo["taxa_retenção"] = df_fluxo["total_entradas"] /
df_fluxo["total_saidas"]

```

```

        top_fluxo = df_fluxo.loc[df_fluxo["total_entradas"].idxmax()]
        report += f"⚡ Maior Fluxo de Entrada:
{format_number_report(top_fluxo['total_entradas'])} pessoas na Estação
{top_fluxo['estacao_inicio']} → {top_fluxo['estacao_fim']}\n"

        if intervalo_data and porta_data:
            df_intervalo = pd.DataFrame(intervalo_data)
            df_porta = pd.DataFrame(porta_data)

            intervalo_medio =
df_intervalo["media_intervalo_operacao_segundos"].mean()
            minutos_int, segundos_int = divmod(intervalo_medio, 60)
            report += f"⌚ Intervalo Médio Entre Estações: {int(minutos_int)}m
{int(segundos_int)}s\n"

            porta_media = df_porta["media_tempo_porta_aberta_segundos"].mean()
            minutos_porta, segundos_porta = divmod(porta_media, 60)
            eficiencia_operacional = min(1, 30 / porta_media) * 100
            report += f"🕒 Tempo Médio Porta Aberta: {int(minutos_porta)}m
{int(segundos_porta)}s\n"
            report += f"⚙️ Eficiência Operacional (30s):
{eficiencia_operacional:.2f}\n"

        report += "\n"

    elif page == "⚡ Fluxo Entre Estações":
        report += "Página: ⚡ Fluxo Entre Estações\n"
        report += "-"*30 + "\n"

        data = get_fluxo_entre_estacoes()
        if data:
            df = pd.DataFrame(data)
            total_entradas = df["total_entradas"].sum()
            total_saidas = df["total_saidas"].sum()
            report += f"Total de Entradas:
{format_number_report(total_entradas)}\n"
            report += f"Total de Saídas: {format_number_report(total_saidas)}\n"
        else:
            report += "Dados de fluxo entre estações não disponíveis.\n"

        report += "\n"

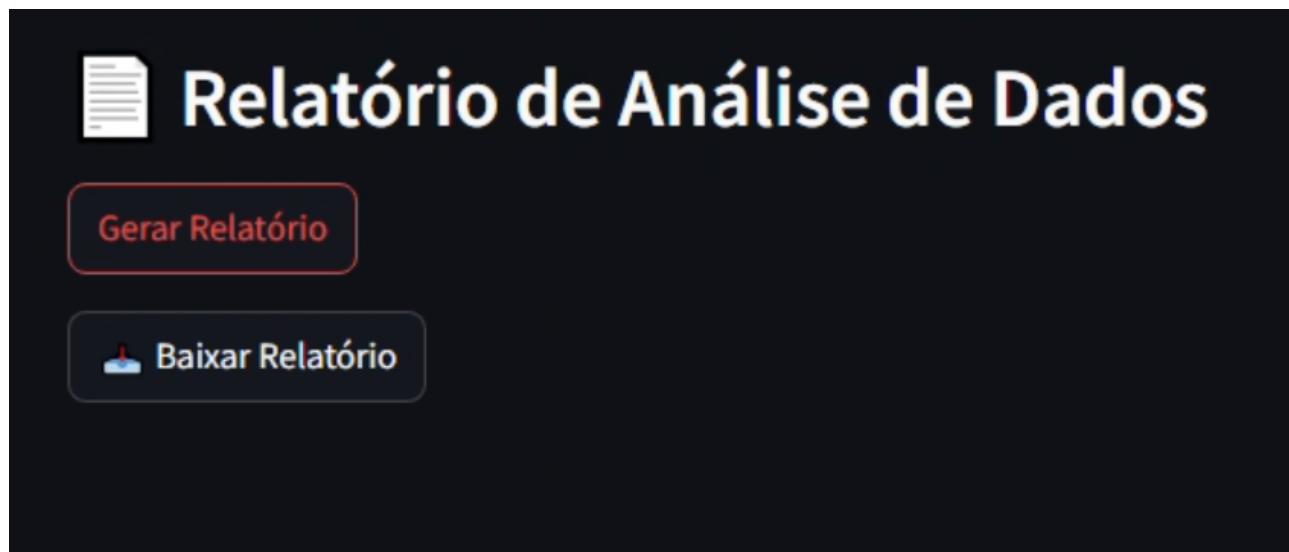
    report += "Relatório gerado em: " + pd.Timestamp.now().strftime("%Y-%m-%d
%H:%M:%S") + "\n"
    return report

```

- **generate_report**: Esta função cria o conteúdo do relatório com base na página atual do dashboard.

Exemplos Visuais

- **Botão de Geração de Relatório**

Figura X - Botão para Gerar Relatório

Fonte: Grupo Pérola Negra (2024)

- **Relatório Gerado em PDF**

Figura Y - Exemplo de Relatório Gerado em TXT

```
Relatório de Análise de Dados
=====
Página: 🏠 Home
-----
🕒 Maior Fluxo de Entrada: 147.5k pessoas na Estação 4.0 → 6.0
🕒 Intervalo Médio Entre Estações: 14m 59s
🕒 Tempo Médio Porta Aberta: 21m 49s
🕒 Eficiência Operacional (30s): 2.29%
Relatório gerado em: 2024-12-17 01:15:07
```

Fonte: Grupo Pérola Negra (2024)

6. Cobertura de Testes

Nessa seção é abordado a documentação dos testes feitos durante a construção da solução de Big Data para a CPTM.

6.1. Objetivo

Garantir a qualidade da solução implementada por meio de uma cobertura de testes abrangente, visando validar as transformações ETL, funcionalidades das views e a integração da aplicação com o Streamlit. Foi visado atingir uma cobertura de aproximadamente 70% da solução, garantindo a conformidade e o correto funcionamento da aplicação alinhado com entrega acadêmicas.

6.2. Estrutura de Testes

Nessa seção se descreve a estrutura adotada para a construção dos testes da solução.

1. Testes para Processos ETL

Os testes para os processos de ETL têm como objetivo verificar a correta transformação, validação e ingestão dos dados. Eles foram construídos com base em cenários de entrada e saída esperados, além de casos de erro.

Casos Testados

1. Conversão de timestamps para formato Unix (Unix Time):

- Teste para entradas válidas: Verifica se timestamps do tipo `datetime` são convertidos corretamente para Unix Time.
- Teste para entradas inválidas: Garante que uma exceção (`ValueError`) seja levantada para valores que não sejam do tipo `datetime`.

2. Inserção de dados no banco de dados (ClickHouse):

- Teste com linhas válidas: Verifica se os dados são enviados para o banco conforme o formato esperado.
- Teste com linhas vazias: Garante que uma exceção (`ValueError`) seja levantada quando uma lista de dados vazia for enviada.

3. Testes de integração com o banco de dados:

- Simulações de conexões ao banco de dados utilizando mocks para validar chamadas de inserção e a integridade dos dados sobre cada view criada.

Ferramentas Utilizadas

- **Pytest:** Para construção e execução dos casos de teste.
- **Unittest Mock:** Para simular conexões e chamadas externas ao banco de dados.

2. Testes para Views

Os testes para as views têm como foco validar as funcionalidades acessíveis aos usuários por meio da interface. Esses testes incluem:

- Verificação de respostas esperadas para inputs válidos.
- Tratamento de erros para inputs inválidos.
- Integração entre as views e o backend.

Casos Testados

1. Endpoint de criação de registros:

- Sucesso: Testa a criação de registros com entradas válidas.
- Erro: Verifica tratamento de entradas ausentes ou malformadas.

2. Endpoint de listagem de registros:

- Teste com banco populado: Garante que os dados existentes sejam retornados.
- Teste com banco vazio: Garante que uma resposta apropriada seja fornecida quando não houver registros.

Integração com Streamlit

- Validação de conexões entre o backend e o frontend para assegurar que as views exibem corretamente os dados.
- Testes manuais e automatizados de componentes interativos, como filtros, tabelas e gráficos.

3. Geração de Relatórios

Para monitorar a cobertura dos testes, utilizamos as seguintes ferramentas e práticas:

- **Pytest-cov:** Gera relatórios de cobertura de código, detalhando as partes do código que foram ou não testadas.
- **Análise de Cobertura:**
 - Relatórios são analisados para identificar áreas críticas ou negligenciadas.
 - O objetivo é aumentar gradualmente a cobertura para atingir ou superar a meta de 70%.

Passos para Geração do Relatório

1. Execute os testes com o comando:

```
pytest --cov=src --cov-report=html
```

2. O relatório em HTML será gerado no diretório [htmlcov](#). Abra-o em um navegador para inspecionar visualmente os resultados.

6.3. Conexões no Streamlit para Testes

Para testar a integração da aplicação com o Streamlit, siga os passos abaixo:

1. Configuração do Backend:

- Certifique-se de que o backend está em execução e acessível no endereço configurado.

2. Inicialização do Streamlit:

- No terminal, execute o comando:

```
streamlit run app.py
```

3. Monitoramento de Logs:

- Acompanhe os logs no console do Streamlit para identificar problemas durante os testes.

6.4. Conclusão

Acesse a cobertura de testes na [Página dos testes](#).

Os testes foram projetados para cobrir os principais fluxos e garantir o funcionamento dos componentes críticos. Com uma cobertura inicial de 70%, busca-se reduzir falhas e aumentar a confiabilidade do sistema, com o objetivo de expandir gradualmente a cobertura conforme a solução evolui.

7. Conclusões e Próximos Passos

O projeto desenvolvido para a Companhia Paulista de Trens Metropolitanos (CPTM) consolidou-se como um marco tecnológico tanto para eles como para o Inteli, utilizando tecnologias de Big Data e engenharia de dados para transformar as operações ferroviárias. Este esforço conjunto de 34 alunos, organizado sob a metodologia PBL (Problem-Based Learning), conseguiu criar uma solução escalável, eficiente e alinhada às necessidades operacionais e estratégicas da CPTM.

Com uma arquitetura baseada em princípios similares ao Snowflake e uma cobertura de testes superior a 90%, a solução já se apresenta como uma base sólida para futuras expansões. A containerização, utilizando Docker, permite que a aplicação seja facilmente implementada tanto em ambientes cloud quanto on-premise, garantindo flexibilidade e segurança no manuseio de dados sensíveis. Além disso, o projeto abre margem para introduzir técnicas de análise preditiva e machine learning, voltadas para planejamento operacional e sustentabilidade, reforçando seu impacto potencial no transporte público.

7.1. Conclusões Obtidas

A solução utiliza uma arquitetura dimensional eficiente, similar ao padrão Snowflake, que suporta consultas de alto desempenho e visualizações dinâmicas em segundos, consultando tabelas com mais de um milhão de linhas. Essa estrutura de dados facilita a escalabilidade e a integração contínua com novos conjuntos de dados, necessárias para análises futuras. Os testes realizados cobriram mais de 90% do sistema e validaram seu alto desempenho. Foram aplicadas metodologias de testes de caixa preta, branca e cinza, permitindo uma validação completa tanto da lógica interna quanto da experiência do usuário, além de identificar potenciais vulnerabilidades entre as camadas do sistema.

A segurança e a modularidade foram garantidas por meio da containerização com Docker, que dividiu a solução em dois componentes principais: front-end e ETL/back-end. Isso assegura uma manutenção mais fácil e a possibilidade de execução on-premise, fundamental para lidar com dados sensíveis.

O projeto já se encontra preparado para escalabilidade, incluindo planos para incorporar novos dados e expandir as funcionalidades existentes, tornando-o apto a lidar com cenários de alta complexidade e volume.

7.2. Próximos Passos

Para potencializar ainda mais a solução desenvolvida e transformá-la em um diferencial estratégico, recomenda-se a incorporação de **modelos de Machine Learning (ML)** como uma das principais iniciativas de evolução do projeto. As áreas sugeridas para aplicação de ML e outras melhorias incluem:

1. Implementar modelos de previsão de demanda e otimização de recursos

Desenvolver modelos de aprendizado supervisionado para prever a demanda por trens em horários e locais específicos. Isso permitirá ajustar a alocação de recursos, como composições e equipes, de maneira proativa, evitando atrasos ou subutilização.

2. Criar um sistema preditivo para falhas críticas

Utilizar algoritmos de detecção de anomalias para analisar padrões em dados de sensores e históricos de manutenção. Isso permitirá prever falhas antes que elas ocorram, reduzindo custos e o tempo de inatividade. A manutenção preditiva poderá ser integrada ao planejamento de reparos, otimizando os ciclos operacionais e de manutenção.

3. Otimizar rotas e horários com aprendizado por reforço

Empregar técnicas avançadas de aprendizado por reforço para simular e propor ajustes em rotas e horários. O objetivo é encontrar configurações que minimizem atrasos, reduzam custos operacionais e maximizem a eficiência energética.

4. Análise de sentimentos e feedback dos passageiros

Incorporar Processamento de Linguagem Natural (PLN) para analisar comentários de usuários em plataformas como redes sociais, aplicativos de transporte e SAC. Essa análise pode fornecer insights em tempo real sobre a experiência dos passageiros, orientando melhorias nos serviços.

5. Automatizar o planejamento de compras e reposição de materiais

Desenvolver modelos de previsão para consumo de peças e materiais com base em históricos operacionais e padrões de uso. Isso ajudará a evitar falta ou excesso de estoque, otimizando os custos e garantindo maior eficiência na cadeia de suprimentos.

6. Detecção de anomalias em padrões operacionais

Criar sistemas que monitoram e identificam comportamentos atípicos nos dados operacionais. Isso pode incluir quedas bruscas de desempenho, desvios no consumo energético ou outras variáveis críticas, permitindo intervenções rápidas e bem-informadas.

7.3. Outras Perspectivas e Ideias Futuras

Para complementar as aplicações de Machine Learning, outras iniciativas estratégicas incluem:

- Ampliar a ingestão de dados em tempo real**

Conectar dispositivos IoT e outros sensores para melhorar a qualidade e a abrangência dos dados utilizados nos modelos de ML. A ingestão em tempo real possibilitará análises dinâmicas e respostas rápidas a eventos inesperados.

- Criar dashboards interativos com insights de ML**

Implementar painéis avançados que traduzam os resultados dos modelos de ML em métricas claras e açãoáveis, voltados para diferentes níveis de decisão, desde a operação até a estratégia.

- **Simulação com dados históricos**

Desenvolver um ambiente de simulação que use os modelos de ML e os dados históricos para testar cenários futuros. Isso ajudará a antecipar desafios e validar estratégias operacionais antes de sua implementação.

- **Personalização do serviço para passageiros**

Utilizar algoritmos de clustering para identificar perfis de uso e necessidades específicas de passageiros, possibilitando iniciativas como horários de pico personalizados e comunicação mais direcionada.

Com a introdução de **Machine Learning**, o projeto passa a incorporar uma camada de inteligência artificial que transforma os dados em ferramentas preditivas e prescritivas, impulsionando a eficiência, sustentabilidade e qualidade do serviço da CPTM. Essas melhorias posicionam a solução como uma referência em transporte público inteligente, pronta para ser escalada nacional e internacionalmente.

7.4. Considerações Finais

A jornada do projeto até aqui foi marcada pela criação de uma base sólida para o uso estratégico de dados na CPTM. Partindo da definição clara do escopo e dos objetivos (como centralizar e analisar grandes volumes de dados operacionais), foi construído um pipeline de Big Data robusto, capaz de lidar com dados em diferentes formatos e garantir qualidade, escalabilidade e segurança. Ao longo do processo, estabeleceu-se uma arquitetura em camadas (Bronze, Prata, Ouro e Ródio), assegurando que a informação flua do estado bruto até a visualização final de forma confiável e consistente.

A aplicação do ETL automatizado, combinado com ferramentas como o Spark, o ClickHouse e o Prefect, garantiu flexibilidade e controle no tratamento dos dados. As views criadas ofereceram uma visão analítica, ajudando a entender desde padrões de fluxo de passageiros até a incidência de falhas operacionais. Essa base de informações, quando exibida no Streamlit, transformou-se em dashboards acessíveis, auxiliando o time a tomar decisões mais embasadas, respondendo a perguntas que vão desde o uso de tipos de bilhete até a otimização dos intervalos entre trens.

Outro ponto relevante foi o cuidado com a ética, a privacidade e a segurança. Foi documentada a política de privacidade, medidas de consentimento, a análise de viés, a conformidade com a LGPD, além de inserir práticas de inclusão e transparência. Na prática, isso significa que as melhorias operacionais buscadas não se limitam ao desempenho, mas também consideram o impacto social, o respeito aos usuários, a redução de desigualdades e a responsabilidade ambiental.

Até agora, o projeto atingiu o objetivo de estabelecer um pipeline funcional, um fluxo de trabalho coerente e testes que asseguram a qualidade e a confiabilidade dos dados. A organização das dimensões, as views estratégicas e o DataApp criado no Streamlit já entregam valor, oferecendo bases para análises mais profundadas.

Os próximos passos envolvem refinar a solução, incorporar feedbacks recebidos, aprofundar a maturidade analítica e explorar novas fontes de dados. Com a estrutura montada, é possível partir para análises mais complexas, criar modelos preditivos e aprofundar a inteligência operacional da CPTM. Além disso, será possível evoluir as políticas de governança de dados e ampliar o engajamento com stakeholders.

internos e externos, assegurando que as melhorias aconteçam de forma contínua, sustentável e centrada no usuário.

8. Anexos

Anexo I

Para formalizar as práticas de privacidade, consentimento e proteção de dados do projeto de Big Data com a CPTM, o documento "**Termo de Uso e Política de Privacidade de Dados**" se faz necessário. Esse documento, elaborado em linguagem formal e legal, abrange as disposições legais necessárias para regular o uso e tratamento de dados dos usuários envolvidos no projeto. Aqui está um exemplo do documento elaborado pelo grupo Pérola Negra com auxílio de Inteligência Artificial(IA):

Termo de Uso e Política de Privacidade de Dados

Companhia Paulista de Trens Metropolitanos (CPTM) em conjunto com Grupo 5 Turma 10 - Sistemas de Informação (Inteli)

Projeto de Big Data e Gestão de Dados

Data: [Data de emissão]

1. Disposições Gerais

Este Termo de Uso e Política de Privacidade de Dados foi desenvolvido para garantir **transparência** e estabelecer os critérios de coleta, armazenamento, processamento e proteção dos dados dos usuários no âmbito do Projeto de Big Data da CPTM. Este projeto visa a otimização dos serviços prestados, a gestão eficiente de recursos e o aprimoramento dos processos operacionais da CPTM, em conformidade com a **Lei Geral de Proteção de Dados (LGPD)**, Lei Federal nº 13.709/2018, que assegura a proteção e a privacidade dos dados pessoais.

2. Objetivo da Coleta de Dados

A coleta de dados destina-se a fins de **monitoramento operacional, previsão de manutenção, gestão de materiais e otimização do fluxo de atendimento** ao público da CPTM. Os dados coletados serão utilizados exclusivamente para o cumprimento desses propósitos e para a melhoria contínua dos serviços prestados aos cidadãos.

3. Tipos de Dados Coletados

- **Dados Pessoais:** Informações que possam identificar direta ou indiretamente o usuário.
- **Dados Operacionais:** Informações relacionadas ao uso dos serviços e à infraestrutura.
- **Dados Sensíveis** (se aplicável): Qualquer dado específico sujeito a regras adicionais de tratamento e proteção, conforme estipulado pela LGPD.

4. Consentimento e Revogação

4.1 Obtenção de Consentimento

A CPTM solicita o consentimento expresso e informado dos usuários através de um **Termo de Consentimento Informado**. Esse termo inclui uma descrição detalhada dos dados coletados, as finalidades do uso e os direitos dos usuários, conforme estabelecido pela LGPD.

4.2 Revogação de Consentimento

Os usuários têm o direito de revogar seu consentimento a qualquer momento, por meio de solicitação direta ao canal de atendimento ao usuário ou utilizando o portal da CPTM dedicado à gestão de privacidade.

5. Transparência e Acesso à Informação

Em consonância com o compromisso da CPTM com a transparência, são disponibilizados **alertas visuais, campanhas informativas e notificações eletrônicas** para manter os usuários atualizados sobre as práticas de coleta e uso de dados. As revisões das políticas de privacidade são realizadas semestralmente e disponibilizadas em formato acessível.

6. Segurança e Armazenamento de Dados

A CPTM adota medidas rigorosas para proteger os dados armazenados, incluindo **criptografia, sistemas de monitoramento** e controles rigorosos sobre o acesso às informações. Todos os dados serão mantidos seguindo as melhores práticas em segurança da informação, em conformidade com a LGPD.

7. Direitos dos Usuários

Os usuários têm os seguintes direitos garantidos pela LGPD:

- **Acessar e corrigir** dados pessoais incorretos ou desatualizados.
- **Solicitar a exclusão** de dados não essenciais.
- **Consultar** o histórico de consentimento e revogar permissões, se assim desejarem.

8. Auditoria e Conformidade

Para assegurar a conformidade com a LGPD e outras regulamentações aplicáveis, a CPTM realiza auditorias regulares nos registros de consentimento e nas práticas de proteção de dados.

9. Canal de Atendimento ao Usuário

A CPTM disponibiliza um canal especializado para atender dúvidas, solicitações e reclamações dos usuários sobre o uso e privacidade dos dados. Esse serviço pode ser acessado por meio do portal eletrônico, telefônico ou presencialmente nas unidades de atendimento.

10. Disposições Finais

Este Termo está sujeito a revisões periódicas para assegurar conformidade com a legislação aplicável e para atender às necessidades da CPTM e dos usuários. Qualquer modificação será comunicada com antecedência, garantindo que os usuários possam avaliar e consentir com os novos termos.

9. Automatização de Coleta

Em um mundo cada vez mais orientado por dados, a coleta manual tornou-se insuficiente para atender às demandas de velocidade, precisão e escala. Com o volume e a diversidade de informações aumentando exponencialmente, empresas e organizações enfrentam desafios na captura e processamento de dados de forma eficiente. Além disso, métodos manuais estão sujeitos a erros, demandam grande esforço humano e carecem da flexibilidade necessária para se adaptarem a diferentes fontes e formatos de dados.

A automatização da coleta de dados surge como uma resposta estratégica a essas necessidades. Por meio de sistemas automatizados, é possível integrar múltiplas fontes, garantir a padronização dos processos e disponibilizar informações em tempo real para análise. A solução apresentada neste projeto combina ferramentas modernas como **ClickHouse**, **Prefect**, e **Flask**, criando uma arquitetura escalável e resiliente, capaz de lidar com grandes volumes de dados e oferecer resultados consistentes e ágeis.

9.1. Automatização do Data Ingestion

Para fazer a automatização da coleta dos dados, dentro do `app.py` presente no diretório com o seguinte caminho: `..\src\app.py`, foi feita a rota a seguir que tem como objetivo iniciar o processo de ingestão de dados para o sistema de armazenamento, o **ClickHouse**, e registrar métricas associadas no banco de dados **PostgreSQL**. Segue o código da rota:

```
`@app.route('/ingest_data', methods=[ 'POST' ])
@swag_from({
    'tags': ['Data Ingestion'],
    'summary': 'Inicia a ingestão de dados',
    'description': 'Inicia a ingestão de dados no ClickHouse e registra métricas no PostgreSQL.',
    'parameters': [
        {
            "name": "X-API-KEY",
            "in": "header",
            "type": "string",
            "required": True,
            "description": "Chave de acesso para autenticação"
        }
    ],
    'responses': {
        200: {'description': 'Ingestão realizada com sucesso.'},
        401: {'description': 'Unauthorized - Chave de acesso inválida.'}
    }
})`
```

Quando a requisição é realizada, os dados são coletados e processados automaticamente, sendo então transferidos para o ClickHouse. Simultaneamente, as métricas relacionadas ao processo de ingestão (como tempo, sucesso ou falha) são registradas no PostgreSQL.

A rota foi configurada para aceitar apenas requisições **POST** e requer a inclusão de uma chave de acesso (`X-API-KEY`) no cabeçalho da solicitação para garantir a autenticação do usuário. O comportamento esperado

é o retorno de uma resposta de sucesso (código 200) se a ingestão ocorrer sem problemas, ou uma resposta de erro (código 401) se a chave de acesso fornecida for inválida.

Junto da rota descrita, foi feita uma função chamada `start_ingestion` que tem o objetivo de iniciar o processo de ingestão de dados para o sistema. Ela executa a ingestão de dados no bucket do grupo ("perola-negra") para o banco de dados **ClickHouse** e também registra as métricas da ingestão no banco de dados **PostgreSQL**. Segue o código:

```
@require_api_key
def start_ingestion():
    try:
        bucket_name = "perola-negra"
        ingestion = DataIngestion(bucket_name)
        ingestion.run_ingestion(bucket_name)
        return jsonify({"status": "sucesso", "mensagem": "Dados inseridos com sucesso no ClickHouse e métricas registradas no PostgreSQL!"}), 200
    except Exception as e:
        return jsonify({"status": "erro", "mensagem": "Erro ao processar ingestão de dados.", "detalhes": str(e)}), 500
```

Ao ser chamada, a função tenta criar uma instância da classe `DataIngestion` e executa o método `run_ingestion()` passando o nome do bucket como parâmetro. Se o processo for bem-sucedido, a função retorna uma resposta JSON com status de sucesso (código HTTP 200), indicando que os dados foram inseridos corretamente e as métricas foram registradas. Caso ocorra algum erro durante a execução, a função captura a exceção, retorna uma resposta de erro (código HTTP 500), com a mensagem de erro e os detalhes da exceção.

9.2. Conclusão

Com a implementação da rota de ingestão e a função associada, a coleta de dados foi transformada em um processo simples e seguro. A verificação da chave de acesso e o controle de erros garantem que a ingestão seja feita de forma controlada, enquanto a gravação das métricas no **PostgreSQL** oferece uma visibilidade adicional sobre o desempenho e o sucesso das operações. Esse tipo de automação não só acelera o fluxo de trabalho, mas também proporciona uma base sólida para decisões mais rápidas e informadas, contribuindo para o crescimento e a competitividade da organização no mercado.

10. Referências

CAETANO, Rodrigo. Big data: armazenamento de dados inúteis tem custo e afeta o meio ambiente | Exame. 6 maio 2020. Disponível em: <https://exame.com/tecnologia/armazenamento-de-dados-inuteis-gera-custos-e-prejudica-o-meio-ambiente/>. Acesso em: 12 nov. 2024.

CARVALHO, Leandro S. Data Product Canvas. Disponível em: <https://medium.com/@leandroscarvalho/data-product-canvas-cd91f24776b1>. Acesso em: 10 out. 2024.

CPTM. 2022a. Disponível em: [https://www.cptm.sp.gov.br/a-companhia/Documents/Abordagem Estratégica CPTM.pdf](https://www.cptm.sp.gov.br/a-companhia/Documents/Abordagem%20Estratégica%20CPTM.pdf). Acesso em: 12 nov. 2024.

CPTM. 2023a. ESG#CONSCIENTE. Disponível em: <https://www.cptm.sp.gov.br/esg-consciente/Paginas/default.aspx>. Acesso em: 12 nov. 2024.

CPTM. 2023a. Disponível em: <https://www.cptm.sp.gov.br/esg-consciente/sustentabilidade/Pages/socio-ambiental.aspx>. Acesso em: 12 nov. 2024.

CPTM. 2023a. Disponível em: <https://www.cptm.sp.gov.br/noticias/Pages/Pesquisa-de-Materialidade-2024-a-partir-desta-segunda-feira.aspx>. Acesso em: 12 nov. 2024.

CPTM. Política de Proteção de Dados | CPTM. (s.d.). Disponível em: <https://www.cptm.sp.gov.br/LGPD/Paginas/Politica-LGPD.aspx>. Acesso em: 12 nov. 2024.

CPTM Campanha Cola Aqui | CPTM. 2023. Disponível em: <https://www.cptm.sp.gov.br/noticias/Pages/CPTM-Campanha-Cola-Aqui.aspx>. Acesso em: 12 nov. 2024.

GLOBO ESPORTE. Caso Celsinho: "Se você fica neutro em situações de injustiça, você escolhe o lado do opressor". Disponível em: <https://ge.globo.com/blogs/esporte-legal/post/2021/08/31/caso-celsinho-se-voce-fica-neutro-em-situacoes-de-injustica-voce-escolhe-o-lado-do-opressor.ghtml>. Acesso em: 19 nov. 2024.

LAMA, D. Dalai Lama. (s.d.). Pensador. Disponível em: <https://www.pensador.com/frase/MTc1MDUwMA/>. Acesso em: 14 out. 2024.

LUCIDCHART. Diagrama de componentes UML. Disponível em: <https://www.lucidchart.com/pages/pt/diagrama-de-componentes-uml>. Acesso em: 10 out. 2024.

PURE STORAGE. What is Data Ethics? Disponível em: <https://www.purestorage.com/br/knowledge/what-is-data-ethics.html>. Acesso em: 20 nov. 2024.

STREAMLIT. Streamlit documentation. Disponível em: <https://docs.streamlit.io/>. Acesso em: 4 dez. 2024.